Author:
Day, James J

Title:
Structural studies of trypanosomal tRNA synthetases as drug targets for African Animal Trypanosomiasis

Structural studies of trypanosomal tRNA synthetases as drug targets for African Animal Trypanosomiasis

James Day

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Biochemistry MScR in the Faculty of Biomedical Sciences. School of Biochemistry, 2018.

Word Count: 13843

Abstract

African Animal Trypanosomiasis is a disease that primarily afflicts cattle in Africa. It affects at least 37 countries in the region and, in 2017, was estimated to be responsible for the overall loss of around £4.75 billion per annum to the rural economy. The various symptoms, ultimately leading to the deaths of over 3 million cattle per year, result in a reduction of every aspect of livestock farming and it is therefore considered a severe repressor of central African development both cultural and economic.

It is caused by the group of parasitic protozoa called *Trypanosoma*, which are carried by all flies of the '*Glossina*' genus (tsetse flies) and renders large swathes of rural Africa – particularly the 'tsetse belt' – unsuitable for cattle farming. The two organisms most significant in causing this widespread economic repression are *Trypanosoma congolense* and *Trypanosome vivax*. Aminoacyl tRNA synthetase enzymes have been previously identified as promising potential drug targets in *T. congolense* and, in this study, that potential is being explored through the generation of crystal structures of those enzymes.

Three structures of *T. congolense* lysyl-tRNA Synthetase I (LysRS1) have been solved following expression in Rosetta 2 pLacI *E. coli* cells and purification using cobalt affinity chromatography and size exclusion chromatography. The structures contain LysRS1 complexed with lysine and AMP (LAM_G4; 1.9 Å), with the lysyl-AMP intermediate (LAM_H2; 2.2 Å) and with the nucleoside analogue cordycepin (LCM_F3; 2.4 Å). These structures provide a detailed description of the *T. congolense* LysRS1 active site and comparing these structures with mammalian LysRS enables identification of differences in the ATP-binding site that could be exploited in structure-based drug-design. Furthermore, conformational changes in a pair of loops associated with the active site may be significant for enzyme activity and specificity.

Dedication and Acknowledgments

I would firstly like to thank my supervisor Professor Leo Brady, and Dr Nick Burton a member of the Brady Group, for their continued expertise, time and assistance during this course. A very big thank you to Storm McCready-Fallon, whose work I have continued over this time and whose help was invaluable in troubleshooting at the beginning. My thanks also go to my progression panel members, Professor Ian Collinson and Professor Wendy Gibson, for their time and helpful feedback.

I would like to thank all the members of C101 for their day-to-day help in the lab and for just generally being great company, especially the technicians who were always very friendly and kept everything running like clockwork. A special thank you to Dr Peter Wilson who was always available to help with anything BioSuite-related – particularly the crystallography robots and imaging hotels.

Finally, big thanks to all my family and friends for their support during this degree.

Author’s Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ..... DATE:.....

Abbreviations

AAT	Animal African trypanosomiasis
Aa-tRNA	Aminoacyl-tRNA
AMP	Adenosine monophosphate
ATP	Adenosine triphosphate
E. coli	*Escherichia coli*
FPLC	Fast protein liquid chromatography
GF	Gel filtration
GlyRS	Glycyl-tRNA synthetase
HAT	Human African trypanosomiasis
IPTG	Isopropyl β-D-1-thiogalactopyranoside
LLG	Log likelihood gain
LysRS	Lysyl-tRNA synthetase
MR	Molecular replacement
MWCO	Molecular weight cut-off
NCI/DTP	National Cancer Institute / Developmental Therapeutics Programme
OD_{500}	Optical density at 500 nm
PEG	Polyethylene glycol
RMSD	Root-mean-square deviation
SASA	Solvent accessible surface area
SDS-PAGE	Sodium dodecyl suphate polyacrylamide gel electrophoresis
SEC	Size exclusion chromatography
Tcon	*Trypanosoma congolense*
T. brucei	*Trypanosoma brucei*
T. congolense	*Trypanosoma congolense*
T. vivax	*Trypanosoma vivax*
UV	Ultraviolet
VDW	Van der Waals

Table of Contents

List of Figures

List of Tables

Introduction

In an individual animal, African Animal Trypanosomiasis (AAT or Nagana, from the Zulu word "N'gana" meaning "powerless/useless") causes symptoms such as anaemia and emaciation, and without treatment can often lead to death. (Muhanguzi et al., 2014) At the herd level it leads to lower fertility rates, stunted growth, lower milk yields and a higher overall mortality rate, especially of infants. From meat and milk production to draft power, Nagana is therefore said to cause a reduction in every aspect of livestock farming such that it can be considered one of the industry's biggest constraints (Holt et al., 2016). Both it and the closely related Human African Trypanosomiasis (HAT) are considered to have historically caused severe repression of Central African cultural and economic development (Steverding, 2008). Due to the close relationship between the Nagana-causing species *Trypanosoma brucei brucei* and its two other subspecies (*Trypanosoma brucei gambiense* and *Trypanosoma brucei rhodesiense*) that cause HAT, there is potential for zoonosis and, in fact, warnings already exist from the Center for Food Security and Public Health in 2009 that non-zoonotic trypanosomes may cause disease in humans with certain genetic defects. Despite all of this, Nagana has seen relatively little pharmaceutical attention when compared to HAT, and so this investigation intends to extend a previous study further exploring the potential of inhibition-based drugs for the treatment of AAT. Ultimately, the hope is that the structures and analyses provided by this study can inform a structure-based approach to drug design.

Trypanosomes

Nagana is the result of infection by members of a group of unicellular parasitic protozoa called *Trypanosoma*, specifically by organisms *Trypanosoma congolense*, *Trypanosoma vivax* and the *brucei* subspecies of *Trypanosoma brucei* (Morrison et al., 2016). The genus was named for its characteristic drill-like motion by David Gruby who, whilst studying a parasitic organism in frog blood in 1843, described its movement as "...like a drill or corkscrew, which is why I propose to name this hematozoan '*Trypanosoma*'." - from the Greek words 'Trypanon' (auger) and 'Soma' (body) (Langousis & Hill, 2014). The large, biting flies that carry them are unique to Africa and feed on the blood of vertebrate mammals. *T. vivax* also exists in South America where several types of biting fly transmit it, in contrast to in Africa where tsetse flies are the exclusive vector (Jones and Dávila, 2001). Figure 1, below, shows the spread of the two species of trypanosomes largely responsible.

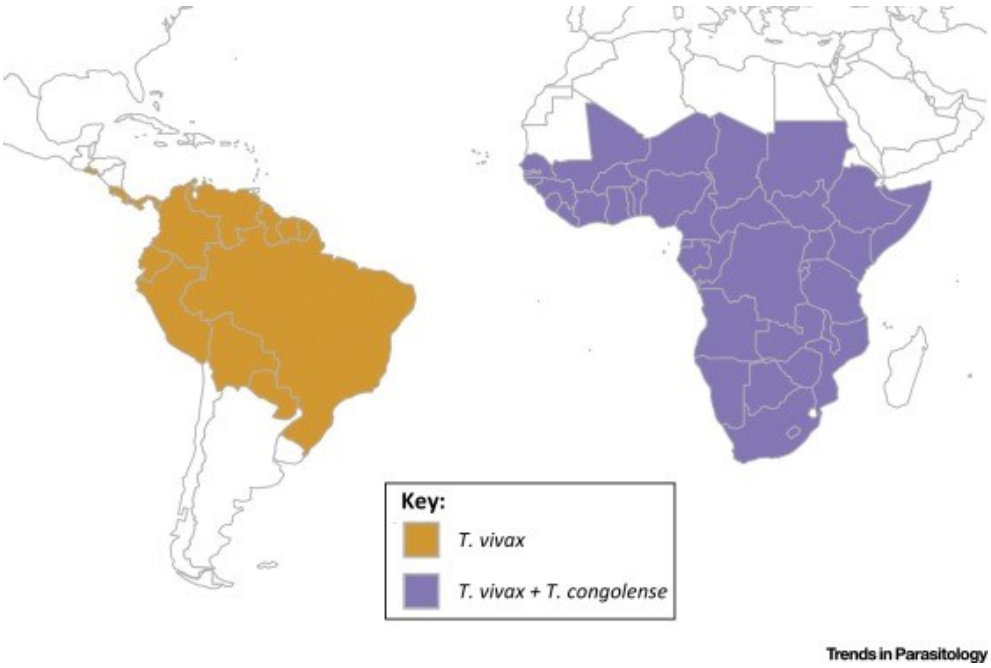

Figure 1: Distribution of *T. vivax* and *T. congolense*.
This image shows the distribution of the two main causes of Nagana (Morrison et al., 2016).

The trypanosome lifecycle can be said to begin when a tsetse fly bites and ingests the blood of an infected mammal, after which the parasite then undergoes a developmental cycle involving proliferation, migration and ultimately differentiation into an infectious form, ripe for transfer to their next mammalian host (Rotureau and Van Den Abbeele, 2013). Many studies simplify these parasites' life cycles as consisting of only two stages – in mammalian blood and in the fly vector – as seen below in Figure 2.

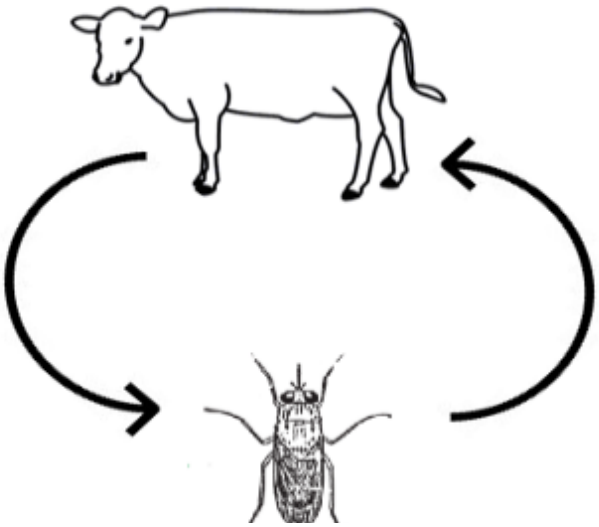

Figure 2: Basic life-cycle of Nagana-causing *Trypanosoma* species.
Simplistically, the Nagana-causing *Trypanosoma* species can be considered to have two stages in their life cycles – in the tsetse fly vector and in the mammalian (cow) host.

However, in reality it is much more complicated: the trypanosomes have a multitude of different morphological forms, corresponding to different stages of their developmental cycle corresponding with different locations within their host or vector (Matthews, 2005). The metabolism of the trypanosome differs depending on its morphological form. For example, energy production is achieved through proline oxidation and oxidative phosphorylation when in the midgut stage compared to a total reliance on glycolysis and substrate-level phosphorylation in the bloodstream form (Eyford et al., 2011). In the fly salivary gland stage, genes are activated that produce a coat of variable surface glycoproteins. This antigen variation mechanism allows trypanosomes to evade neutralisation by the immune system of the mammalian host and, when the mammalian bloodstream form is ingested by a tsetse fly, these genes are inactivated when the parasite returns to the fly midgut stage (Horn, 2014). All of these changes contribute to the parasite being near-impossible to 'pin down' by the host immune system. Instead, the hope is to target a process present and necessary at all stages of the life cycle: protein synthesis.

Three distinct developmental pathways have been defined, based on how each species of trypanosome travels within the alimentary tract of the fly (Rotureau and Van Den Abbeele, 2013). *T. vivax* has a developmental pathway that lasts around three days and is contained within the proboscis and cibarium of the fly. Members of the *Nannomonas* subgenus, e.g. *T. congolense*, develop as they pass through the midgut, foregut and proboscis (Gibson, 2013). All types of *T. brucei*, including those that infect humans, have a developmental cycle that can last up to three weeks as they migrate through the midgut, foregut, proboscis and salivary glands. This is summarised in Figure 3 below.

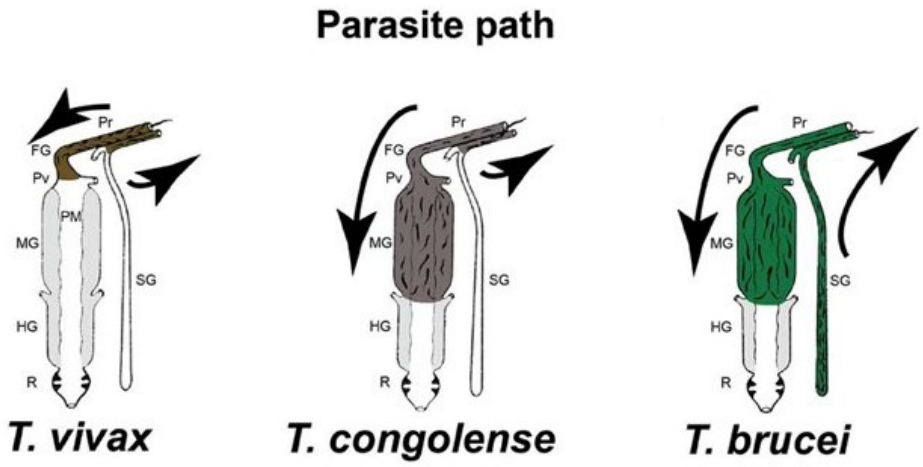

Figure 3: *Trypanosoma* in tsetse flies.
This briefly summarises the movement of the three different types of nagana-causing *Trypanosoma* species within the tsetse fly vector. At each location, the *Trypanosoma* species assume a different morphological form.
Pr: proboscis, FG: foregut, Pv: proventriculus, PM: peritrophic matrix, MG: midgut, HG: hindgut, R: rectum, SG: salivary glands (Rotureau and Van Den Abbeele, 2013).

As members of the kinetoplastea class, trypanosomes have unusual 'kinetoplast' structures: large networks of circular DNA called kDNA that are found within the mitochondria of species belonging to this class (Lukes et al., 2002). These structures, along with the presence of other unusual features such as the glycosome organelle that compartmentalises glycolysis (Michels et al., 2006), distinguish a vast evolutionary distance between the intracellular mechanisms of trypanosomes and that of their mammalian hosts. This may make it easier to develop specific compounds to inhibit the growth of the parasites, with as little interference as possible to the ordinary cells of the host, by targeting proteins that reflect these evolutionary differences.

Aminoacyl-tRNA synthetases

When considering potential avenues into drug development for diseases caused by microbial organisms, protein synthesis stands out as a process that has been thoroughly validated as an excellent target for disruption (Orelle et al., 2013). There already exists a wide range of antimicrobial drugs that function by interfering with protein synthesis at a variety of different stages (Wilson, 2009). Protein synthesis is made doubly attractive as a drug target in parasites because they are often dependent on enhanced rates of protein synthesis compared to their hosts, in order to support their rapid growth, expansion and development within the host. It is therefore reasonable to assume that they are particularly sensitive to having these rapid rates reduced or even stopped.

Aminoacyl tRNA synthetases have already been shown to be essential for the *in vitro* survival and

growth of *T. brucei* (Kalidas et al., 2014) – a close relative of *T. congolense* and *T. vivax* – and are already established as drug targets in bacterial species (Nakama et al., 2001). They are ancient enzymes that attach an amino acid to its corresponding tRNA through the catalysis of an esterification reaction between a specific amino acid, or its precursor, and one of its compatible cognate tRNAs to form the 'aminoacyl tRNA' (aa-tRNA) that acts as a substrate in RNA translation. This process is referred to as 'charging' or 'loading' and is shown below as Figure 4.

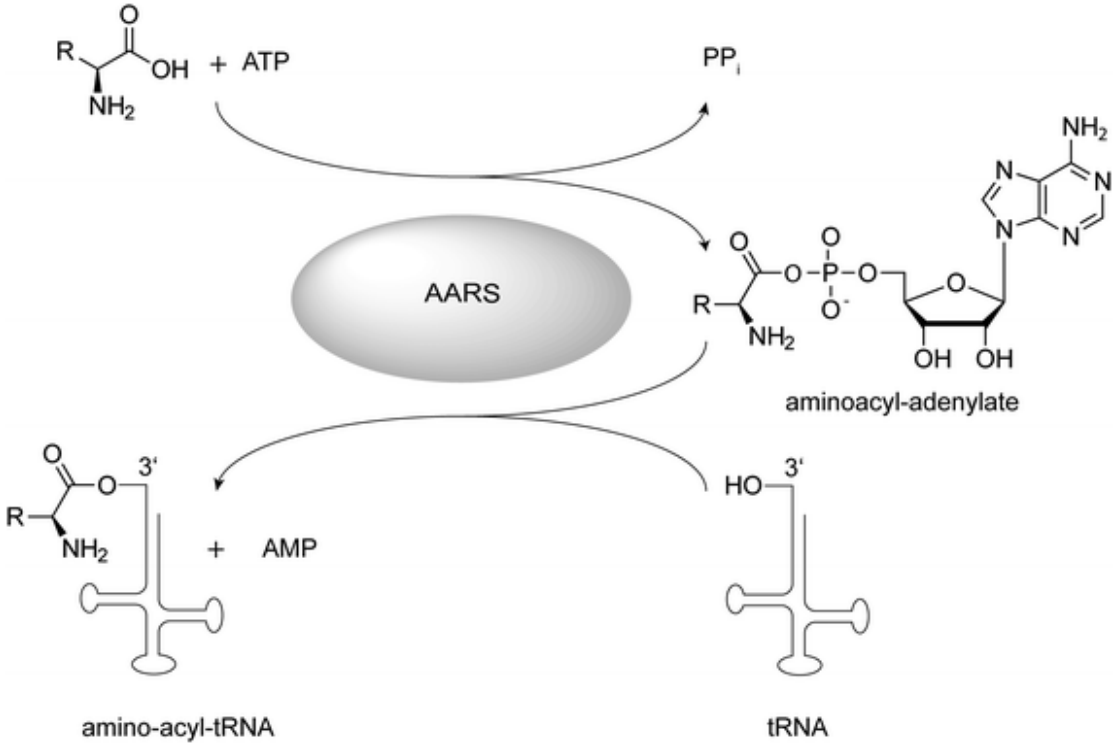

Figure 4: Aminoacyl-tRNA synthetase primary mechanism.
The appropriate amino acid binds alongside ATP to the enzyme and, with the energy provided by the release of inorganic pyrophosphate, an aminoacyl-adenylate is formed. The tRNA then binds to the enzyme and is joined to the amino acid via an esterification reaction, releasing AMP (Battenberg et al., 2013).

Essentially, this is a two-step reaction that can be written as follows:

$$AA + AaRS + ATP \rightleftharpoons AaRS * AA - AMP + PPi \ (1)$$

$$AARS * Aa - AMP + tRNA \rightleftharpoons \mathrm{Aa - tRNA + AMP + AaRS} \ (2)$$

After the charging is complete, a ribosome can then transfer the amino acid from the aa-tRNA onto a growing peptide as dictated by the genetic code. Aa-tRNA synthetase, in facilitating the formation of aa-tRNA, therefore plays an integral role in protein translation. Although this is their main role, they have evolved over millennia to host a variety of additional functions in signalling pathways and

other metabolic processes (Freist, 1995). There are twenty-four different aminoacyl-tRNA synthetases in *T. congolense*, with some amino acids being charged by several different types of the enzyme. For example, there are thought to separate genes for mitochondrial and cytoplasmic tryptophanyl-tRNA synthetase, as identified by the presence (or absence) of a mitochondrial signalling sequence (McCready-Fallon, 2018).

Structurally, these are multi-domain proteins containing both a catalytic domain and an anticodon domain. The catalytic domain facilitates the 'charging' process whereas the anticodon domain interacts with the anticodon region on the RNA to ensure the amino acid is being affixed to the correct tRNA (Rogers et al., 1994). Incorrect pairings usually involve two amino acids that have similar shapes but different properties. Some aminoacyl tRNA synthetases contain a hydrolytic 'editing' domain, enabling them to cleave incorrect couplings by hydrolysis of the ester linkage. (Beuning and Musier-Forsyth, 2000) On the tRNA, the esterification occurs at the hydroxyl group of the terminal adenosine nucleotide. There are two classes of aminoacyl tRNA synthetase: class I, which aminoacylates at the 2'-hydroxyl and has two highly conserved sequence motifs, and class II which aminoacylates at the 3'-hydroxyl and has three highly conserved sequence motifs (Eriani et al., 1990) Members of the former are typically monomeric or dimeric, whereas the latter are typically dimeric or tetrameric. These classes are believed to have evolved from different ancestors via convergent evolution (Pouplana and Schimmel, 2001).

Lysyl-tRNA synthetase has been identified as a class II synthetase in trypanosomes and comes in two forms: LysRS1 and LysRS2. Although it has not been experimentally verified in any trypanosomes, LysRS1 has an N-terminal extension that, in humans, shows chemokine activity. LysRS2 has a C-terminal extension domain that is, in *T. brucei*, involved in mitochondrial localisation. It is for this reason that LysRS2 is believed to be the kinetoplast form, meaning LysRS1 should be the cytoplasmic form (McCready-Fallon, 2018).

X-Ray Crystallography

When discovered in 1895 by W.C. Röntgen, the mysteriously named 'X-rays' were subject to great public interest. Intense research into their qualities and potential applications began almost immediately in multiple directions (Tubiana, 1996). As early as 1896, physicians were using X-rays on patients to image parts of the skeleton and specific organs and in 1897, when J.J. Thomson showed that X-rays were able to ionise gas, further study of this quality led to the discovery of electrons. In 1913, X-ray crystallography was developed by W.H. Bragg and his son W.L. Bragg. It involves the interpretation of the diffraction pattern created when the crystallised form of a pure molecular sample is exposed to X-rays (Bragg and Bragg, 1913). Originally limited to simple molecules such as salts, the types of molecule on which the technique could be used grew steadily over the first half of the twentieth century. The first crystal structure of a biological macromolecule

– that of sperm-whale myoglobin – was published by John Kendrew in 1957 with a resolution of 6 Å, further improved to 2 Å by 1960. In 1980, the double-helix structure of DNA was solved through this method (Wing et al., 1980) although it had previously been inferred via fibre diffraction studies (also with X-rays). As of February 2018, there have been over 142,000 depositions to the protein data bank, the vast majority of which are protein structures determined using X-ray crystallography (Berman et al., 2000). This highlights how exponentially the use of this technique has grown. It is currently the most commonly used experimental method to obtain a detailed model of a large molecule.

A particle is said to have been scattered when it collides with another particle (called the scatterer) and the direction of its motion is altered. In a light microscope, an image is compiled by rays of light being coherently scattered by the features of an object and entering a lens, which then reconstructs an image of the object on the opposite side of the lens. For scattering to be significant enough to be detected, the wavelength of the light must be of similar dimension to the object being imaged. Visible light (400-700 nm wavelength) thus cannot produce an image of individual atoms in protein molecules, where bonded atoms are only about 0.15 nm apart. Electromagnetic radiation of this wavelength falls into the X-ray range; X-rays vary in wavelength from 0.01 to 10 nm and therefore are scattered by (the electrons from) even the smallest of molecules. (Drenth, 2007) Hydrogen atoms in proteins, however, are still hard to resolve – instead, knowledge of bond angles, bond lengths and conformational restraints in small molecules can be used to deduce their positions, under the assumption that they behave in proteins as they do in those molecules (Rhodes, 2000).

Although individual atoms scatter X-rays, two factors would ordinarily prevent the production of a focused image of a single molecule. First, current lenses are not capable of focusing X-rays. Secondly, most X-rays will pass through a single molecule without colliding with it. The pattern from that single molecule is too weak to be detected as it has been produced by the scattering of too few photons. Instead, crystals (which contain many millions of identically orientated molecules in an ordered array) are used. A regular array of waves is produced when an X-ray beam interferes with a regular array of scatterers, and so the crystal acts as an amplifier for the scattering of X-rays by the molecules (Wlodawer et al., 2008).

Although this produces stronger, more easily detectable X-rays, the crystal imposes restrictions on the scattering pattern and so the original continuous scattering pattern is now subject to the phenomenon of diffraction. This means that the pattern, formerly a continuous blur of varied intensities over the detector film, is now sampled only at discrete angular positions (‘spots’) on the film where Bragg’s law ($n\lambda=2d\sin\theta$, where n is an integer, d is the distance between atomic layers in the crystal and θ is the X-ray angle of incidence) is fulfilled: where constructive interference is at

its strongest (Bragg and Bragg, 1913). These spots, called reflections, make up a three-dimensional lattice on the detector film called the diffraction pattern. The intensity of each reflection must be measured, and their index assigned, and their relative phases then derived - the X-ray detector can record the intensities, but not the phases, of the X-rays. The phase of an X-ray contains information that is vital to the determination of electron density distribution in the crystal and so the loss of this crucial data during the process is known as the ‘phase problem’ (Taylor, 2003). Deriving the relative phases of the reflections is a major challenge for protein crystal structure solution and is usually achieved either via experimental phasing (most commonly using the anomalous scattering of X-rays by certain atoms) or, as in this study, by molecular replacement. This involves deriving initial estimates of phases from existing, similar protein structures aligned within the unit cell of the unknown structure (Evans and McCoy, 2008). These calculations, and the subsequent refinement of the model parameters and resulting phases, are performed computationally using extensive program suites developed within the crystallography community.

The main aim of this investigation was to establish an understanding of the structure of the *T. congolense* lysyl-tRNA synthetase enzyme, including its complexes with its usual ligands and with potential inhibitors. This was to be achieved using protein X-ray crystallography, and therefore initially required the preparation and purification of highly pure recombinant forms of the enzyme suitable for crystal formation. Although the project initially focused solely on ‘LysRS1’, the cytosolic form of the lysyl tRNA synthetase, there was also scope to extend these studies to other trypanosomal tRNA synthetase enzymes if time permitted.

Materials & Methods

Expression and Purification

The *T. congolense* lysyl-tRNA synthetase I (LysRS1) gene had already been cloned into the hexa His-tagged vector pOPINE (Addgene plasmid #26043) during a previous study (McCready-Fallon, 2018). The LysRS1 sequence within that vector is shown below in Figure 5.

MELSYYDTRIAMVKEMGPLGAAYPHKFHRDHTIPAFRELFKPMLKDKGQRLDKVVTIAGR
IVVKRSSGSKLHFLALQGDGEVLQVISSSGDYVGDGFTDIHSKIKRGDIIGVRGVPSLSNTGE
FSMSAQEITLLSTCYHMLPDEHYGLASVEQRFRQRYLDLIVNRENAKTFITRSKIISYIRSFFD
NRDFLEVETPVLNQIAGGAAARPFITHHNELNQKMYLRIAPELYLKELVVGGLDRVYEIGK
QFRNEGIDLTHNPEFTSVESYWAYADYNDWMETTEELLHGLVMHLYGTPFVKYAPKDSE
GQQVPEVTFNFTRPFKRIQIVPRLEEILGLKFPDEFESVEANAFFLDTCKRNKVECNPPHTTT
RLLDALISHYLEPECHDPTFLCDHPRVMSPLAKWHRKDLRLSERFELFINKKEICNAYTELN
SPLVQREEFEKQLRDREKGDDEAMSIDEGYVQALEYALPPTGGWGLGIDRLVMYLTSQNN
IKEVLLFPAMKPEGMTSVSYPPGTMLNGQGVPLLKHHHHHH-

Figure 5: LysRS1 protein sequence.
The pOPINE hexa His-tag is highlighted in red.

The vector was transformed into competent cells (Rosetta 2 pLacI – Novagen) as per the manufacturer's instructions. Transformants, after being supplemented with 34 mg/mL chloramphenicol, 50 mg/mL carbenicillin and 1% w/v glucose, were then used to produce an overnight culture that was incubated overnight at 30°C whilst shaking at 180 rpm (Innova 44 Incubator shaker series). This was then diluted by a factor of 100 and used to produce 1 L of liquid culture each, supplemented with the same additives as the overnight culture. Liquid cultures were grown in a 37°C incubator whilst shaking at 200 rpm. When they surpassed an OD_{600} reading of 0.55 (measured using a DeNovix DS-11 spectrophotometer), they were transferred to a 20°C incubator for a final thirty minutes and then had 1 mM IPTG added to each culture to induce expression. At this point they were incubated overnight at the same temperature whilst shaking at 180 rpm. The cultures were then spun down in a centrifuge at 5000 *g* for 45 minutes (Sorvall Evolution RC), after which the supernatant was discarded, and the cell pellets transferred to a new container (in most cases a Falcon tube).

Pellets were resuspended in a volume of cobalt load buffer (25 mM HEPES pH 7.5, 500 mM NaCl, 5% w/v glycerol, 45 mM Imidazole) equal to $1/100^{th}$ of the original volume of liquid culture, with 1 protease inhibitor tablet added per 10 mL of mixture. The mixture was processed using a cell homogeniser to ensure a uniform consistency and then passed through a cell press (Constant Systems Cell Disruptor T5) at 25 kpsi to lyse the constituent cells. The resulting lysate was centrifuged at 18000 rpm for 45 mins (Sorvall RC6) and the supernatant was retained, whilst the pellet (cell debris) was discarded.

Using a peristaltic pump (Cole-Parmer minipump variable flow), the supernatant was then loaded

onto a 5 mL cobalt FPLC column (Generon) and washed through using cobalt load buffer. The column was set up on an ÄKTA afterwards and the protein was subsequently eluted using cobalt elution buffer (25 mM HEPES pH 7.5, 500 mM NaCl, 5% w/v glycerol, 500 mM imidazole) on a linear gradient increasing from 0-60% over 30 mins. Promising fractions were identified using the UV trace on the ÄKTA and samples from each were taken and stored for later analysis with SDS-PAGE. The promising fractions were then pooled together and concentrated using a 30,000 MWCO concentrator tube (Generon Vivaspin 20) in a centrifuge (Sorvall ST40R) to a volume less than 5 mL. Using an ÄKTA, the concentrated sample was loaded onto a SEC column (Generon 6-600 kDa HR resin) and washed through using GF buffer (25 mM HEPES pH 7.5, 500 mM NaCl, 5% w/v glycerol). Just as before, the UV trace on the ÄKTA was used to monitor protein elution and samples from each promising fraction were taken and stored for later SDS-PAGE analysis. The His-tagged protein was predicted as ~61.2 kDa by the ExPASy ProtParam webtool (Gasteiger et al., 2005). The promising fractions were then, again, pooled and concentrated down to a suitable volume. Protein concentration was measured using the Protein280 application on a nanospectrophotometer device (DeNovix DS-11 spectrophotometer).

When the cobalt FPLC column degraded in quality due to frequent usage, it was recharged through the following steps. First it was stripped by washing through with 5-10 column volumes of stripping buffer, washed with at least 5-10 column volumes of cobalt load buffer and then with 5-10 column volumes of distilled water. This was all conducted using the peristaltic pump. To recharge the now-washed column, 2.5 mL of 0.1 M cobalt chloride was loaded onto the column. Then it was washed through with 5 column volumes of distilled water and 5 column volumes of cobalt load buffer. Finally, it was stored in 20% ethanol.

Crystallisation

The purified protein, at a known concentration, could then be used in crystallisation trials. In the meantime, it was stored at 4°C. If additives (such as L-lysine, ATP, $MgCl_2$ or an inhibitor) were to be added, the solution would be made up with final concentrations in mind and then kept on ice for around fifteen minutes whilst the crystallisation apparatus was set up. All crystallisation attempts used the vapour diffusion method.

Most crystallisation experiments took place in 2-drop 96-well MRC crystallisation plates, from Molecular Dimensions, using purchased matrix screens (JCSG-plus and PACT Premier). From successful results a custom matrix screen was devised and, additionally, 24-well plates were used early in the project to try and optimise crystallisation conditions in a similar manner.

To conduct the experiments, several protocols were used on a Phoenix nanolitre pipetting robot (Art Robbins) to automate the dispensing of 50 µL of screen solution to plate reservoirs, 0.5 µL of

screen solution to the plate sitting drops, and ultimately 0.5 µL of sample solution to the sitting drop as well (thus a total drop size of 1 µL). Sealed plates were incubated at both 4°C and 20°C in a crystal hotel and examined regularly using Formulatrix imaging software under both visible and cross-polarised light.

Crystallisation was attempted with a wide range of conditions – with both JCSG-plus and PACT premier screens, at both 4°C and 20°C, with different concentrations of protein (5-20 mg/mL) and with different combinations of additives at different concentrations (e.g. 100 mM L-lysine, 10 mM ATP, 10 mM $MgCl_2$), and also without any additives.

When promising crystals were produced, they were transferred at first to 1 µl drops of cryoprotectant solution using nylon loops. Later, it was found that cryoprotection was unnecessary and so they were – from then on – transferred directly into a storage puck within liquid nitrogen. The puck was then loaded into a dewar of liquid nitrogen, after which it was taken to Diamond Light Source for diffraction.

Diffraction data collection and processing

Diffraction data were collected on the Diamond light Source I04-1 beamline after crystals were loaded directly from the dewar into the cryostream via robotic arm. Initially, three test images of each crystal were taken (45° apart) to determine the potential quality of diffraction. For those deemed promising, a full 1800 images were collected with an oscillation of 0.2°, 0.1 second exposure time and at a wavelength of 0.9159 Å. The detector was set to allow collection of data with a maximum resolution of 2.50 Å.

iMosfilm (Battye et al., 2011) was usually used to process raw diffraction data. This program automatically identifies diffraction peaks, then indexes and integrates the data to produce a file containing reflection intensities and indices (mtz format) that could then be scaled and merged using Aimless. Aimless, and other programs mentioned below unless stated otherwise, are part of the CCP4 suite (Winn et al., 2011). In some cases, mtz files were taken directly from the auto-processing functions (Autoproc & Xia2) that are automatically performed on the Diamond ISPyB website. This was done if the auto-processed data already looked usable – determined by the statistics provided on ISPyB such as whether the correct space group was selected, the completeness was good, and the resolution was useful.

The Matthews cell analysis program was first performed to estimate the number of molecules in the asymmetric unit. PHASER, a molecular-replacement program that uses maximum-likelihood methods, was used to produce initial estimates of the phases by placing a search model within the asymmetric unit through both rotational and translational searches after first correcting for

anisotropy. An existing *T. congolense* LysRS1 structure (McCready-Fallon, 2018) served as the model for this search after all waters and ligands had been removed. The rotational and translational search results were filtered based on atomic clashes (packing analysis) and then simple rigid-body refinement was performed on the placed molecules. PHASER then calculates the theoretical phases (α_{calc}) of the model placed in the crystal unit cell and an initial electron density map is then generated through using these phase estimates in conjunction with the amplitudes observed in the actual experiment with the unknown crystal (Evans & McCoy, 2008).

The resulting model then underwent a series of refinement cycles. Each involved adjusting the arrangement of atoms within the model molecule against the electron density map using the program Coot, and then optimising the fit using the program Refmac. As the model is adjusted the phase estimates derived from the model improve (i.e. better fit the crystal structure), leading to improved electron density maps (2Fobs-Fcalc) which, in turn, are used to further adjust the model. This iterative process was used to improve the accuracy of the model structure and was monitored by assessing the fit between model and map via R_{calc} and R_{free} values. The online web-tool Molprobity (Chen et al., 2010) was used to check that the stereochemistry parameters of the model, e.g. Ramachandran values and rotamer orientations, were consistent with the combined observations found in well-refined prior protein crystal structures. Amendments were made with the Molprobity results in mind, particularly when outliers were detected, though in some areas the electron density contradicted the Molprobity evaluations; in these instances, the electron density took priority. Existing structures in the PDB were sometimes used as a guide yet, just as with the Molprobity results, priority was always given to the experimentally derived electron density data. For example, with the attempt at solving an ATP-bound structure, a human lysyl-tRNA synthetase with ATP bound (3BJU – Guo et al., 2008) was used as a comparative reference for active site location and ligand orientation. The ligands were added only after the actual protein chain was deemed satisfactory, to avoid biasing the data.

In addition to the tools provided by CCP4, the Phenix suite (Adams, 2010) was sometimes used as well. For example, when a structure first reached the point where Refmac5 made only negligible improvements, ‘AutoBuild’ was sometimes used to re-start the refinement. This software combines the density modification and chain tracing functions from the RESOLVE program with the automated refinement program phenix.refine to generate a model. It runs several parallel processes that each prioritise a different quality-indicator, and so progress in slightly different ways, and then amalgamates the results of each into a final model. Following any AutoBuild run, refinement would then continue as before using Refmac5 with consideration given to similar structures and Molprobity evaluations. When the structures approached a level that could be considered satisfactory, and Refmac5 again began to return only negligible improvements, phenix.refine was used in the final few iterations. Finally, the online wwPDB validation service was

used to provide an overview of final structure quality.

For some comparisons made between the structures produced in this investigation and those of LysRS enzymes from other species reported previously, BLASTp (Altschul, 1990) was used to assess the similarity between sequences. AreaIMol (Lee and Richards, 1971), a program within the CCP4 suite, was used to calculate the solvent accessible surface area in different structures – as well as how much of it is buried during the binding of certain ligands. The structures were visualised in 3D using PyMOL (DeLano, 2002), which also predicted hydrogen bonding and allowed for structure alignments. The program LigPlot+ (Laskowski and Swindells, 2011) was used to predict hydrophobic and VDW contacts between ligand and binding site, and to produce supporting figures visualising these contacts alongside hydrogen bonds shown more clearly.

Inhibitors

To select potential inhibitors for use in this study, the NCI/DTP database was searched for molecules with over 80% similarity to the lysyl-AMP intermediate. Cladosporin was not included in this study due to sourcing difficulties; it is too expensive to buy from a commercial source and cannot be easily synthesised. The considered inhibitors are shown below in Table 1.
Of the acquired inhibitors (structures shown in Figure 6), cordycepin and clofarabine were trialled in co-crystallisation with LysRS1. This was accomplished in the same way as the ATP trials, only that ATP was substituted with the inhibitor in question. Although the magnesium ions are thought to assist in phosphate positioning, and these nucleoside analogues have no phosphates, magnesium was still included in case it had other effects on the shape of the active site that could affect binding propensity.

Potential inhibitors
Cordycepin
2-fluorocordycepin
Fludarabine
Vidarabine
Cladribine
Clofarabine
Adenine xyloside

Table 1: Seven potential inhibitors of LysRS1 from the NCI database with over 80% similarity to the lysyl-AMP intermediate.
Those acquired are highlighted in green.

Figure 6: Acquired inhibitors.
This shows the structure of the three inhibitors acquired for this study: cordycepin (left), cladribine (middle) and clofarabine (right).

Results & Discussion

Expression and Purification

The protein was successfully produced, extracted and purified through the several processes.

A representative cobalt FPLC trace from the ÄKTA is seen in Figure 7. The well-defined peak shows that this stage is working as intended; the unwanted contents have been washed off and then, as the gradient of elution buffer is increased, the protein is gradually eluted to produce promising fractions (A7-B7 in this instance).

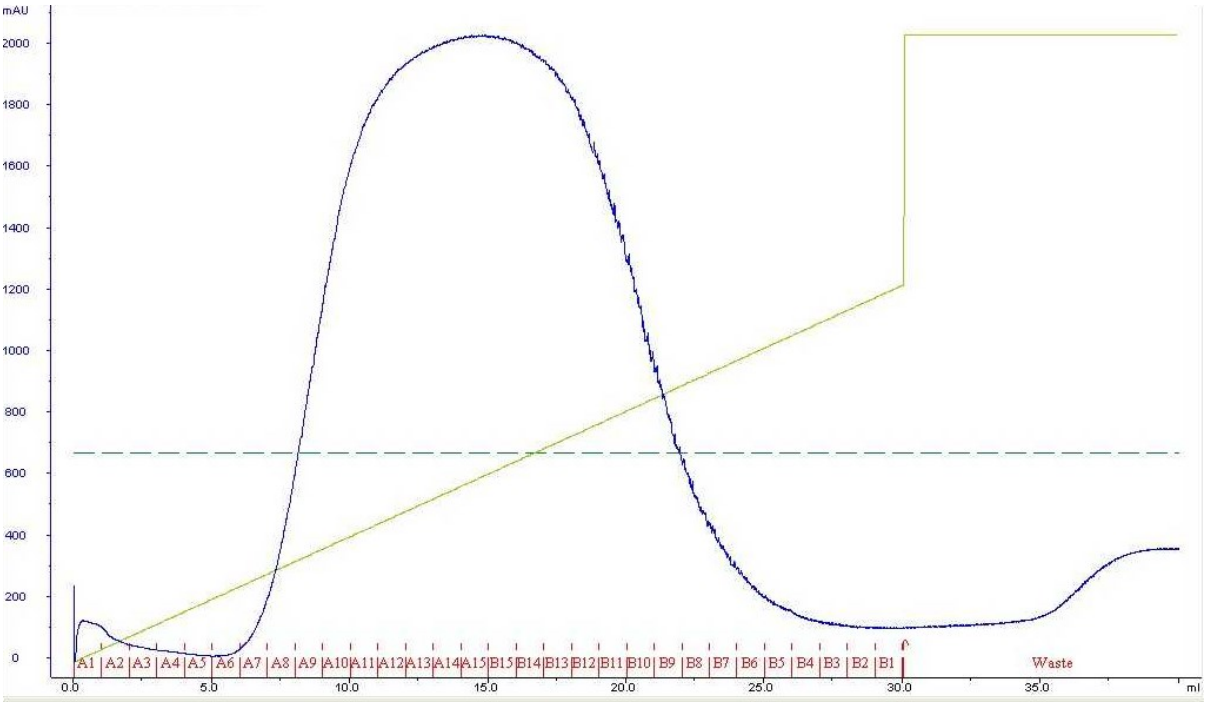

Figure 7: Cobalt FPLC column trace graph.
This is an example trace graph tracking the progress of a cobalt FPLC column run. The blue line measures UV absorbance, whereas the dark green line indicates the elution buffer gradient. Fractions are labelled at the bottom in red.

Figure 8 below shows one of the SDS-PAGE gels run – it displays the effects of the cobalt FPLC column step and provides evidence that the correct protein is being expressed. The thick banding in the promising fraction lanes is situated around 70 kDa despite the molecular weight of the His-tagged LysRS1 protein being predicted as ~61.2 kDa. A similar discrepancy occurred in a previous study (McCready-Fallon, 2018) where peptide mass fingerprinting was then used to verify that the bands do contain LysRS1. Many proteins run at different molecular weights than predicted values on a gel, often due to them running as elongated rather than globular structures. This discrepancy has therefore been ignored. The protein is known to dimerise in solution via a dimeric interface, rather than any disulphide bonding, hence why it is seen as monomeric on the denaturing SDS-PAGE gel.

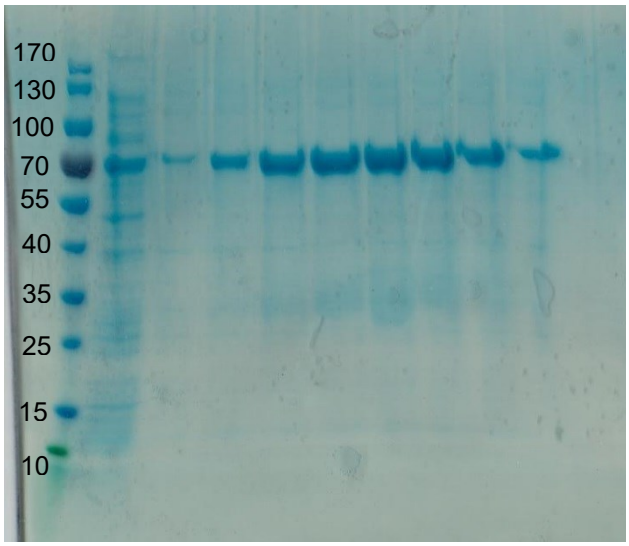

Figure 8: Example Post-Cobalt SDS-PAGE gel.
From the left, the first lane contains the marker. The second shows the contents of the sample before the column. The remaining lanes show the contents of promising fractions.

Figure 9 displays another representative trace from the SEC; another clear peak, with the rest of the UV trace remaining comparatively flat, indicates that the protein is being washed through the column successfully whereas the few remaining impurities portrayed in the post-cobalt FPLC gel are not.

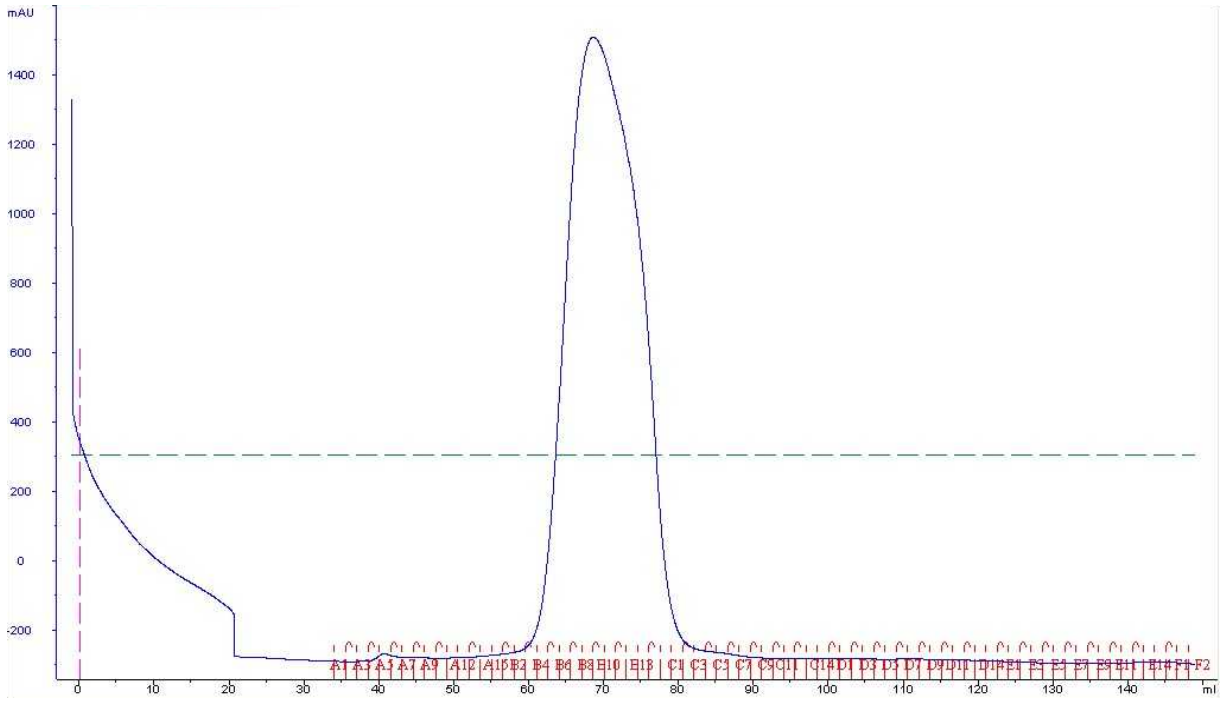

Figure 9: SEC column trace graph.
An example trace graph tracking the progress of a SEC column run. The blue line measures UV and the fractions are labelled at the bottom in red.

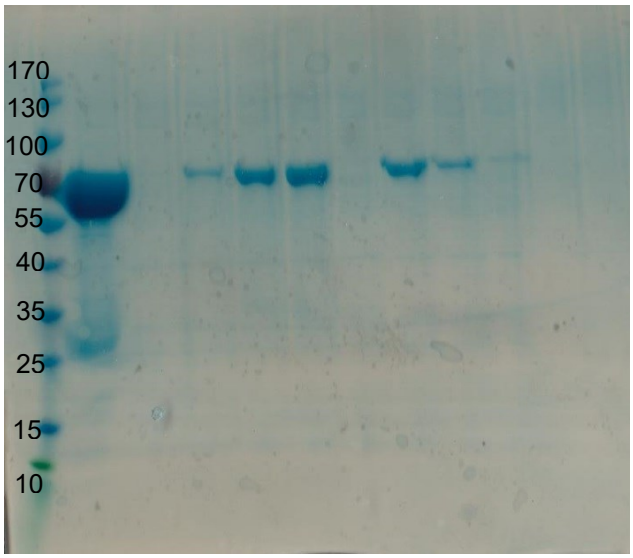

Figure 10: Example Post-SEC SDS-PAGE gel.
From the left, the first lane contains the marker. The second shows the contents of the sample before SEC. The remaining lanes show the contents of promising fractions.

Figure 10 shows another of the SDS-PAGE gels, this one run after a SEC column, showing that the protein has been further purified by this step and confirming that the protein has not been lost during the process.

Crystallisation

LysRS1 could be crystallised from a variety of different well conditions, though there was no success when using a custom screen based on the most frequently successful wells of the commercial 96-well PACT premier screen. Early attempts at optimising crystallisation conditions in 24-well plates were unsuccessful, potentially due to difficulties in storing plates at 4°C; all crystals obtained in this investigation have been grown at 4°C, with no success at 20°C.

Three structures have been produced in this investigation, each containing different ligands bound to the LysRS1 structure. Structure ‘LAM_H2’ has the lysyl-AMP intermediate bound to one chain and the other only containing lysine. Structure ‘LCM_F3’ has the nucleoside analogue cordycepin bound in both chains alongside lysine. Structure ‘LAM_G4’ contains AMP and lysine in the active sites of both chains. Of these, the most important are the former two – LAM_H2 is thought to be a good representation of the enzyme structure during activity and LCM_F3 provides an opportunity to evaluate how a nucleoside analogue could function as an inhibitor to this system. That said, LAM_G4 has the highest resolution and provides supporting information to the two other structures.

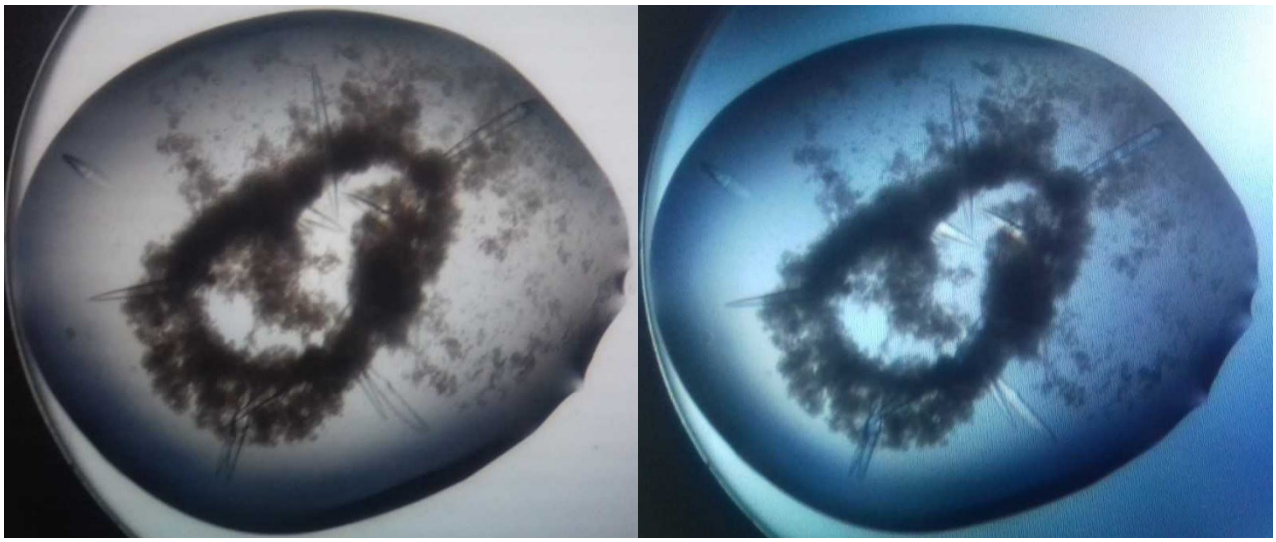

Figure 11: LysRS1 + L-Lysine + ATP + MgCl2 (LAM_G4) crystals.
Sample solution consisted of 20 mg/mL protein, 100 mM L-lysine, 10 mM ATP and 10 mM $MgCl_2$. 0.5 µl of this was added to 0.5 µl well (G4, PACT Premier) solution – 20% PEG 3350, 0.2 M KSCN and 0.1 M Bis-Tris Prop pH 7.5 buffer. Incubated at 4°C and inspected regularly.

The crystals that diffracted to give LAM_G4, the highest resolution LAM (L-lysine + ATP + $MgCl_2$) structure, were produced in a 96-well plate using a commercial PACT premier screen, within the well 'G4' that contained 20% w/v PEG 3350, 0.2 M KSCN and 0.1 M Bis-Tris Prop pH 7.5 buffer. The protein solution itself consisted of 20 mg/mL protein, 100 mM L-lysine, 10 mM ATP and 10 mM $MgCl_2$ – high additive concentrations were used to maximise the chance of binding. They formed within twelve hours and are shown in Figure 11.

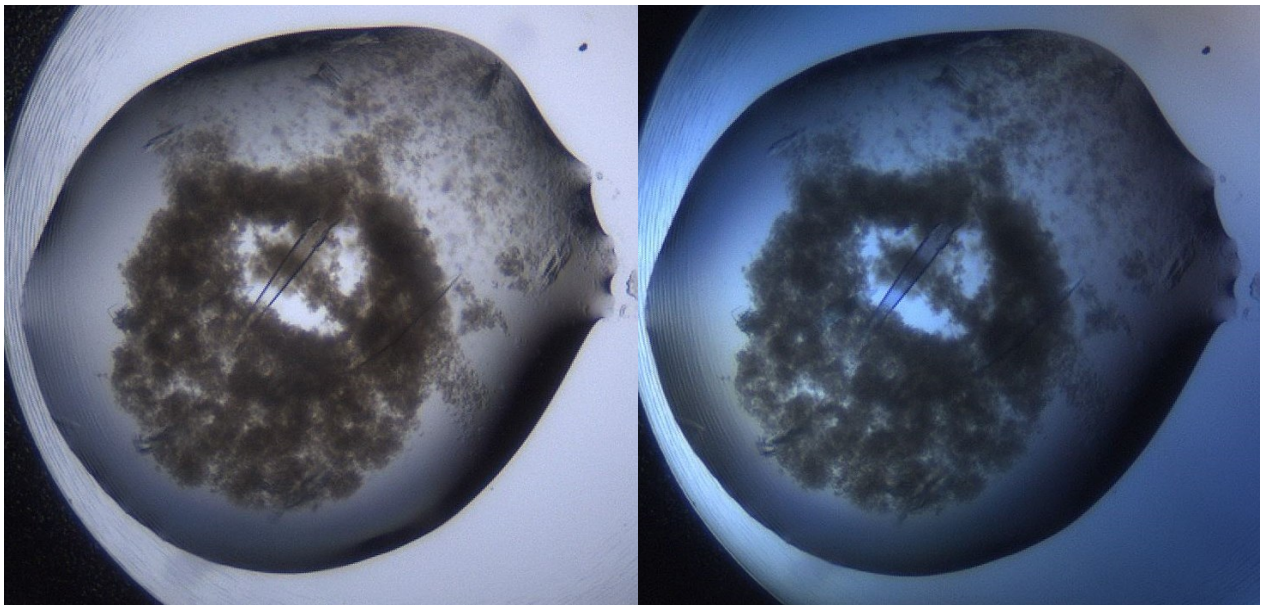

Figure 12: LysRS1 + L-Lysine + ATP + $MgCl_2$ (LAM_H2) crystals.
Sample solution consisted of 20 mg/mL protein, 100 mM L-lysine, 10 mM ATP and 10 mM $MgCl_2$. 0.5 µl of this was added to 0.5 µl well (H2, PACT Premier) solution – 20% PEG 3350, 0.2 M NaBr and 0.1 M Bis-Tris Prop pH 8.5 buffer. Incubated at 4°C and inspected regularly.

The LAM_H2 crystals were also produced in a 96-well plate using a commercial PACT premier screen, within the well ‘H2’ that contained 20% PEG 3350, 0.2 M NaBr and 0.1 M Bis-Tris Prop pH 8.5 buffer. This protein solution was the same as in LAM_G4 – 20 mg/mL protein, 100 mM L-lysine, 10 mM ATP and 10 mM $MgCl_2$. They are shown in Figure 12.

Finally, LCM_F3 came from crystals produced in well ‘F3’ in a 96-well plate using a commercial PACT premier screen. This well contained 20% PEG 3350, 0.2 M NaI and 0.1 M Bis-Tris Prop pH 6.5 buffer. The protein solution consisted of 20 mg/mL protein,100 mM L-lysine, 10 mM cordycepin and 10 mM $MgCl_2$. These crystals are shown in Figure 13.

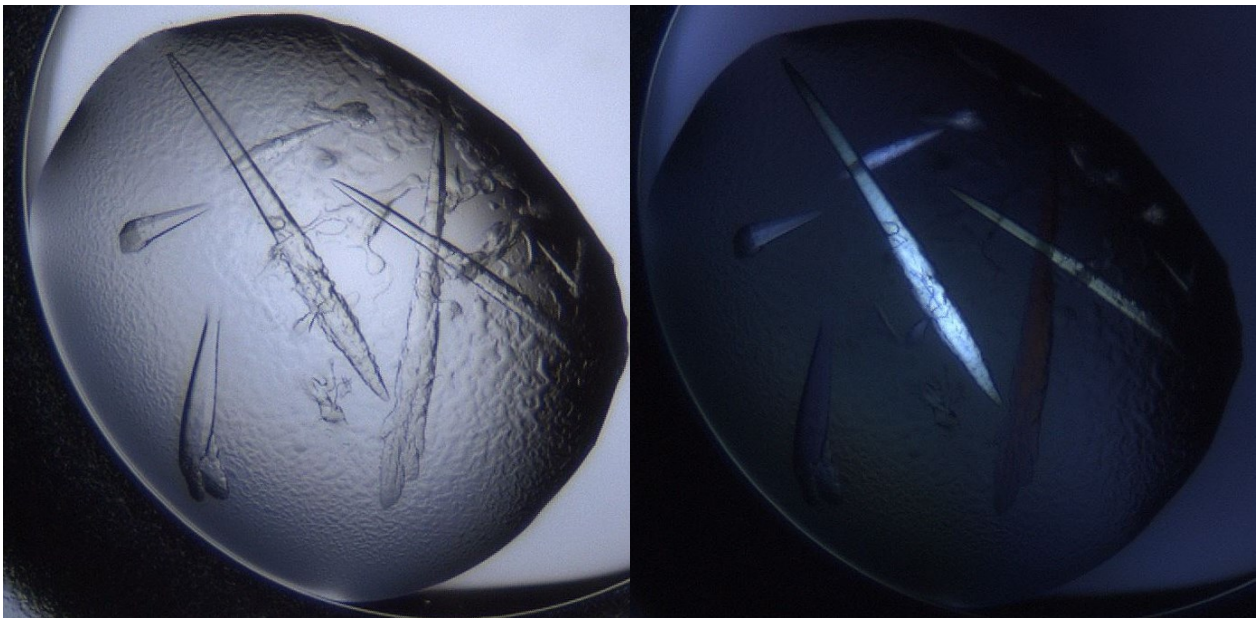

Figure 13: LysRS1 + L-Lysine + Cordycepin + $MgCl_2$ (LCM_F3) crystals.
Sample solution consisted of 20 mg/mL protein, 100 mM L-lysine, 10 mM Cordycepin and 10 mM $MgCl_2$. 0.5 µl of this was added to 0.5 µl well (F3, PACT Premier) solution – 20% PEG 3350, 0.2 M NaI and 0.1 M Bis-Tris Prop pH 6.5 buffer. Incubated at 4°C and inspected regularly.

Diffraction Data Collection and Processing

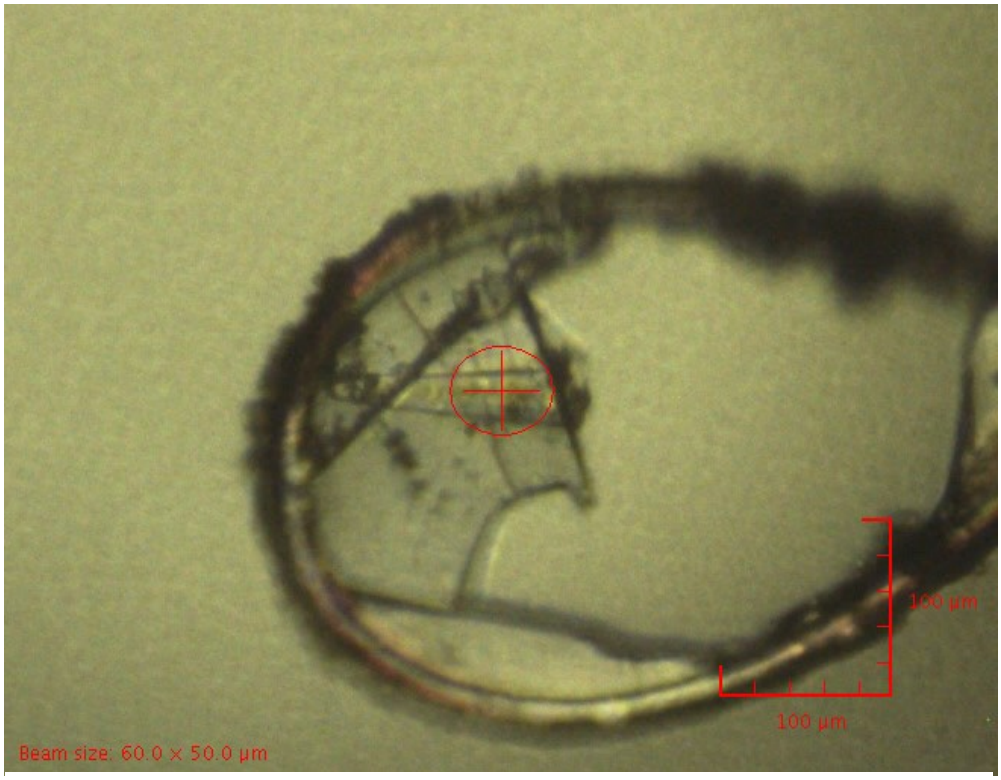

Figure 14: LysRS1 LAM_G4 crystal.
The crystal is pictured within a nylon loop at Diamond Light Source. It was found to belong to the space group $P6_522$ and diffracted to 1.9Å

With a previous batch of crystals, it was found that cryoprotection was unnecessary. Therefore, the crystals that produced the best structures had been taken and had diffraction data collected at Diamond Light Source without cryopreservation. Figure 14 shows the LAM_G4 crystal within the loop on the I04-1 beamline and Table 2 shows the resulting data collection statistics for this crystal.

Data collection LAM_G4	
Resolution range (Å):	95.6-1.9 (2.0-1.9)
Number of reflections [unique]:	92586 [4421]
Wavelength (Å):	0.9159
Unit cell edges a, b, c (Å):	110.3, 110.3, 377.6
Unit cell angles α, β, γ (°):	90.0, 90.0, 120.0
Space group:	$P6_522$
Multiplicity:	27.9 (25.8)
Completeness (%):	98.0 (100)
I/sig(I):	9.1 (1.5)
R_{merge}:	0.5 (4.7)

Table 2: Data collection statistics for LAM_G4 crystal.

Data collection LAM_H2	
Resolution range (Å):	89.2-2.2 (2.3-2.2)
Number of reflections [unique]:	97547 [3127]
Wavelength (Å):	0.9159
Unit cell edges a, b, c (Å):	103.0, 103.0, 406.9
Unit cell angles α, β, γ (°):	90.0, 90.0, 120.0
Space group:	$P6_122$
Multiplicity:	33.9 (31.2)
Completeness (%):	100 (100)
I/sig(I):	10.3 (0.9)
R_{merge}:	0.1 (6.8)

Table 3: Data collection statistics for LAM_H2 crystal.

Data collection LCM_F3	
Resolution range (Å):	96.1-2.4 (2.5-2.4)
Number of reflections [unique]:	103239 [2550]
Wavelength (Å):	0.9159
Unit cell edges a, b, c (Å):	111.0, 111.0, 378.9
Unit cell angles α, β, γ (°):	90.0, 90.0, 120.0
Space group:	$P6_122$
Multiplicity:	40.5 (38.5)
Completeness (%):	99.5 (99.7)
I/sig(I):	7.0 (0.7)
R_{merge}:	0.2 (15.0)

Table 4: Data collection statistics for LCM_F3 crystal.

The initial data collection statistics are seen above in tables 2, 3 and 4. The Matthews cell analysis program in CCP4 was used to estimate the number of molecules in the asymmetric unit. The results from the use of this program are summarised in tables 5, 6 and 7. They each show that there is a very high probability of two molecules in the asymmetric units.

Nmol/asym	Matthews Coeff	%solvent	P(1.94)	P(tot)
1	5.03	75.54	0.00	0.01
2	2.51	51.09	0.99	0.98
3	1.68	26.63	0.01	0.01
4	1.26	2.18	0.00	0.00

Table 5: Matthews program result for LAM_G4.
The table indicates the different potential solutions and then the probability for each solution, calculated using the high resolution limit (P(1.94) in this case) and the total probability across all resolution ranges P(tot).

Nmol/asym	Matthews Coeff	%solvent	P(2.22)	P(tot)
1	5.09	75.83	0.00	0.01
2	2.54	51.67	0.98	0.98
3	1.70	27.50	0.01	0.01
4	1.27	3.34	0.00	0.00

Table 6: Matthews program result for LAM_H2.
The table indicates the different potential solutions and then the probability for each solution, calculated using the high resolution limit (P(2.22) in this case) and the total probability across all resolution ranges P(tot).

Nmol/asym	Matthews Coeff	%solvent	P(2.44)	P(tot)
1	5.51	77.67	0.01	0.01
2	2.75	55.35	0.87	0.86
3	1.84	33.02	0.12	0.13
4	1.38	10.69	0.00	0.00

Table 7: Matthews program result for LCM_F3.
The table indicates the different potential solutions and then the probability for each solution, calculated using the high resolution limit (P(2.44) in this case) and the total probability across all resolution ranges P(tot).

Tables 8, 9 and 10 show summaries from the PHASER molecular replacement results for each structure. These tables detail the best potential solutions for the translational and rotational searches undertaken by the program. LLG is a measure of quality that refers to how much better the data can be predicted using the found model compared to using a random distribution of the same atoms. The Z-score indicates how many standard deviations above the mean that potential solution is. The best solution was clear in each instance as shown by the LLG and Z-score of said solution being notably higher than the others.

Number	LLG	Z-Score
$$ loggraph $$		
1	42.42	8.72
2	28.05	6.70
3	27.47	6.62
$$		

Number	LLG	Z-Score
$$ loggraph $$		
1	349.32	25.92
2	321.68	24.23
3	284.45	21.39
4	121.51	14.38
5	116.45	14.58
6	108.48	13.18
7	100.97	13.19
8	100.06	13.16
9	94.69	13.11
$$		

Table 8: PHASER MR rotational (left) and translational (right) searches for LAM_G4.
The tables detail the top potential solutions ranked by log likelihood gain (LLG).
and then the scores for each.

Number	LLG	Z-Score
$$ loggraph $$		
1	21.05	5.17
2	17.72	4.75
3	11.94	4.02
4	9.23	3.68
5	7.81	3.50
6	7.00	3.40
7	6.74	3.37
8	6.49	3.33
9	5.79	3.24
10	5.78	3.24
11	5.76	3.24
12	5.11	3.16
13	4.71	3.11
$$		

Number	LLG	Z-Score
$$ loggraph $$		
1	137.77	16.17
2	79.95	11.15
3	40.61	10.93
4	34.58	10.77
5	32.23	10.90
6	29.89	10.26
7	27.75	10.45
8	23.39	9.83
9	21.46	10.20
10	20.85	9.92
11	19.17	9.74
12	18.39	10.34
13	17.82	9.59
14	17.70	10.25
15	13.94	9.82
16	13.24	9.97
17	11.49	9.82
18	10.52	9.54
19	9.61	9.40
20	9.57	9.36

Table 9: PHASER MR rotational (left) and translational (right) searches for LAM_H2.
The tables detail the best potential solutions ranked by log likelihood gain (LLG). The top 20 possible solutions of the translational search are shown.

Number	LLG	Z-Score
$$ loggraph $$		
1	22.52	5.50
2	16.95	4.78
3	10.81	3.99
4	10.06	3.89
5	8.92	3.74
6	8.05	3.63
7	7.89	3.61
8	7.41	3.55
9	7.39	3.54
10	7.26	3.53
11	7.24	3.52
12	7.16	3.51
13	6.59	3.44
14	6.48	3.43
15	6.17	3.39
16	6.15	3.38
17	6.06	3.37
18	6.03	3.37
19	6.01	3.37
20	5.98	3.36

Number	LLG	Z-Score
$$ loggraph $$		
1	131.42	15.36
2	118.89	14.21
3	31.90	9.98
4	24.35	10.15
5	17.75	9.09
6	15.00	9.05
7	12.44	9.17
8	8.24	8.82
9	8.13	8.77
10	6.65	8.63
11	5.77	8.96
12	3.81	9.45
13	3.16	9.44
14	2.87	9.23
15	2.63	9.18
16	2.25	8.66
17	1.90	8.53
18	0.44	8.22
19	0.23	8.80
20	0.07	8.79

Table 10: PHASER MR rotational (left) and translational (right) searches for LCM_F3. The tables detail the top potential solutions ranked by log likelihood gain (LLG). In each case the top 20 possible solutions are shown.

In all structures the final translational function Z-score (LAM_G4, 114.3; LAM_H2, 107.1; LCM_F3, 90.0) exceeded 8 indicating, according to PHASER guidelines, there was a very high probability this was a correct solution.

All three structures have been subject to numerous refinement cycles to improve the models and to try and determine whether there are bound ligands and, if so, what they are. The statistics of the latest iterations are shown below in tables 11, 12 and 13.

Refinement statistics (LAM_G4)	
Resolution range (Å):	95.6-1.9 (2.0-1.9)
R_{work}/R_{free} (%):	20.4/24.6
Protein atoms:	8048
Water atoms:	1008
Ligand atoms:	102
Protein B-factor (Å^2):	34.6
Water B-factor (Å^2):	37.3
Ligand B-factor (Å^2):	70.3
R.m.s. deviations	
Bond lengths (Å):	0.003
Bond angles (°):	0.704

Table 11: Refinement statistics for LAM_G4 structure.

Refinement statistics (LAM_H2)	
Resolution range (Å):	89.2-2.2 (2.3-2.2)
R_{work}/R_{free} (%):	19.0/23.9
Protein atoms:	8051
Water atoms:	888
Ligand atoms:	88
Protein B-factor (Å^2):	43.3
Water B-factor (Å^2):	44.0
Ligand B-factor (Å^2):	59.2
R.m.s. deviations	
Bond lengths (Å):	0.003
Bond angles (°):	0.684

Table 12: Refinement statistics for LAM_H2 structure.

Refinement statistics (LCM_F3)	
Resolution range (Å):	96.1-2.4 (2.5-2.4)
R_{work}/R_{free} (%):	19.6/25.6
Protein atoms:	8066
Water atoms:	683
Ligand atoms:	56
Protein B-factor (Å^2):	53.5
Water B-factor (Å^2):	49.8
Ligand B-factor (Å^2):	60.4
R.m.s. deviations	
Bond lengths (Å):	0.005
Bond angles (°):	0.735

Table 13: Refinement statistics for LCM_F3 structure.

As seen in Table 13, the structure LCM_F3 has an average water B-factor lower than that of the protein chain itself. This is likely the result of a more conservative approach to modelling waters in which only tightly bound waters have been included.

Overall structure

The overall fold of the structures solved in this investigation is consistent with previous reports of class II aminoacyl-tRNA synthetases (Ling Joy Pang, 2014) and as previously described for this same enzyme (McCready-Fallon, 2018). In brief, each LysRS1 molecule consists of two main components: the anticodon-binding domain and the aminoacylation domain. The anticodon binding domain is constructed around a beta-barrel at the N-terminus whereas the aminoacylation domain

is comprised of seven antiparallel beta-strands flanked by a series of alpha helices of varying length, an arrangement seen very rarely in non-aaRS proteins.

Each structure contains two molecules of LysRS1 forming a back-to-back dimer. Motif 1 is found at the dimeric interface between each monomer, whereas motifs 2 and 3 are part of the aminoacylation site. These structural features are highlighted below, shown as a monomer in Figure 15 for simplicity and then in its observed dimeric form in Figure 16.

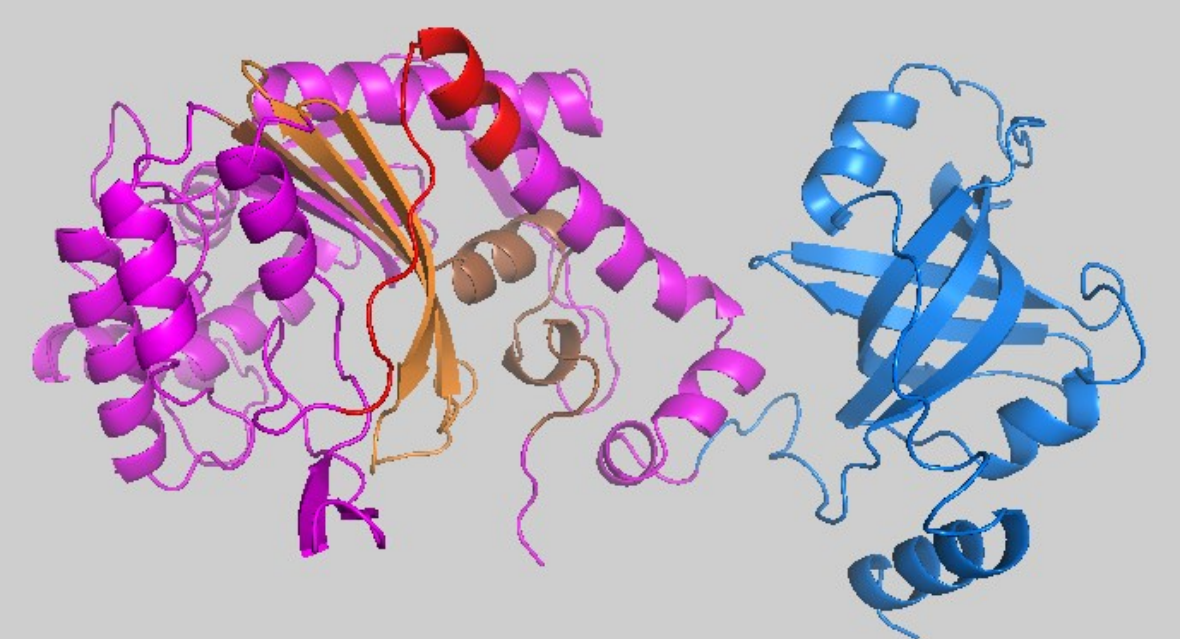

Figure 15: Overall structural view of LAM_H2 monomer.
The N-terminal anticodon binding domain is coloured blue and the C-terminal aminoacylation domain is coloured magenta. Within the aminoacylation domain, motifs 1, 2 and 3 are coloured red, orange and brown, respectively. This image was produced using chain B of LAM_H2.

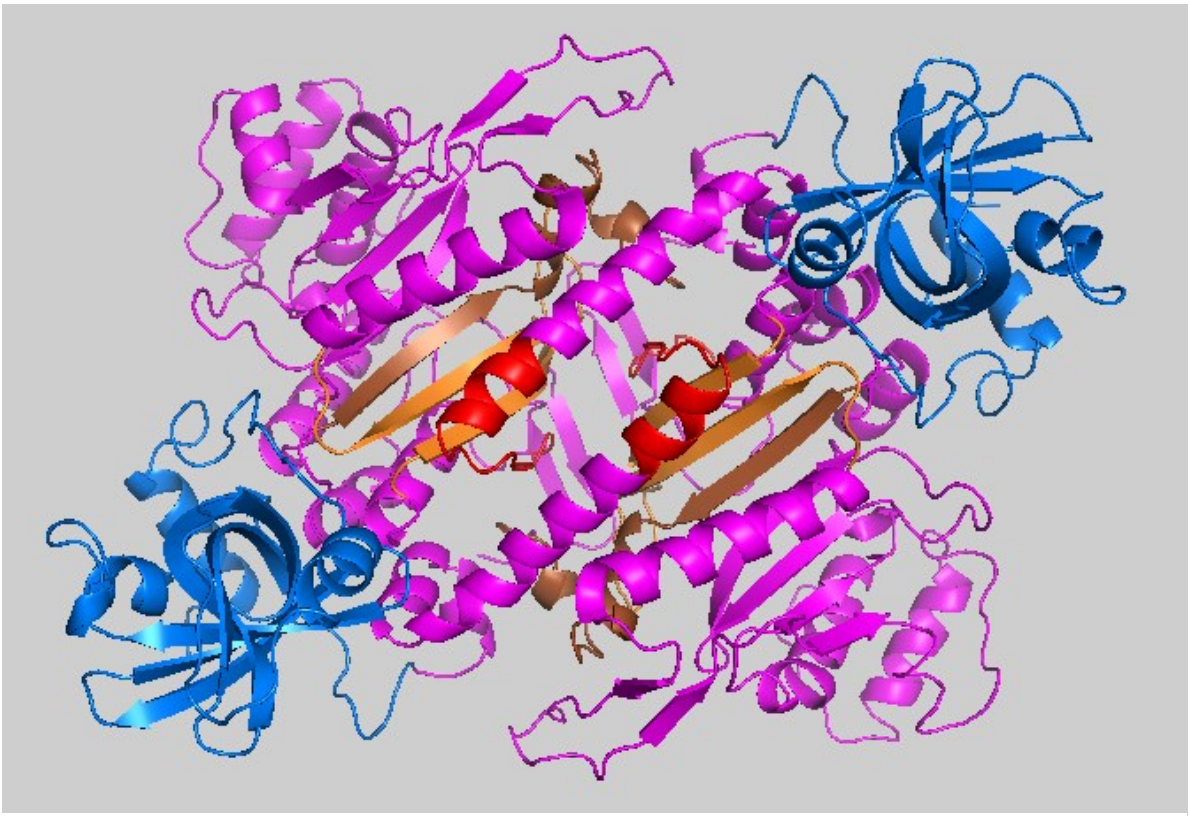

Figure 16: Overall structural view of LAM_H2 dimer.
The N-terminal anticodon binding domain is coloured blue and the C-terminal aminoacylation domain is coloured magenta. Within the aminoacylation domain, motifs 1, 2 and 3 are coloured red, orange and brown, respectively.

These figures were produced using the LAM_H2 structure but there are no major differences in the structural elements of the three structures solved in this investigation. To show this, C-alpha traces have been produced for each structure and then overlaid against the existing Tcon LysRS1 structure. Figure 17 displays this coloured by structure.

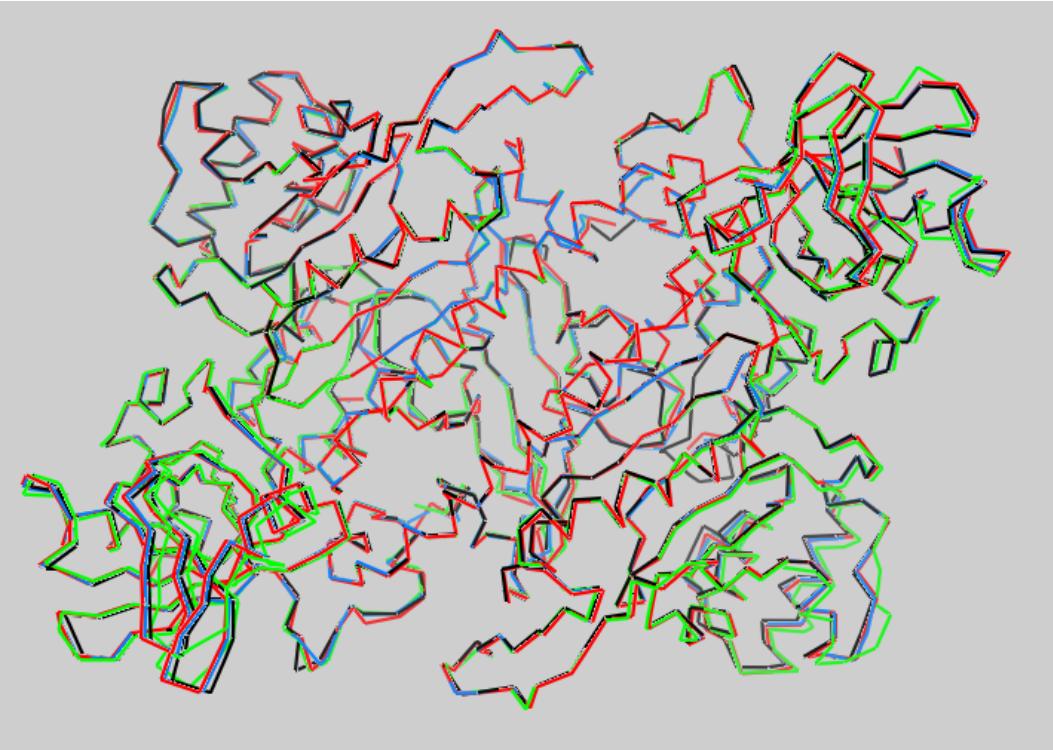

Figure 17: Combined C-alpha trace overlay.
This figure shows the C-alpha traces of four structures overlaid: the three structures from this investigation (LAM_H2; green, LAM_G4; blue and LCM_F3; red) and the existing Tcon LysRS structure fully-complexed with lysine (black).

In order to highlight any major differences between these structures and the existing one, another trace has been prepared that shows the same comparison as above but this time coloured by RMSD. As expected, Figure 18 shows that these structures are very similar except for two loops that will be examined later in more detail. In addition to those loops, there are some areas where the RMSD values are higher. These are thought to be insignificant due to being located on external parts of the protein where the chain is more flexible and (as a consequence of this) the electron density maps are less clear.

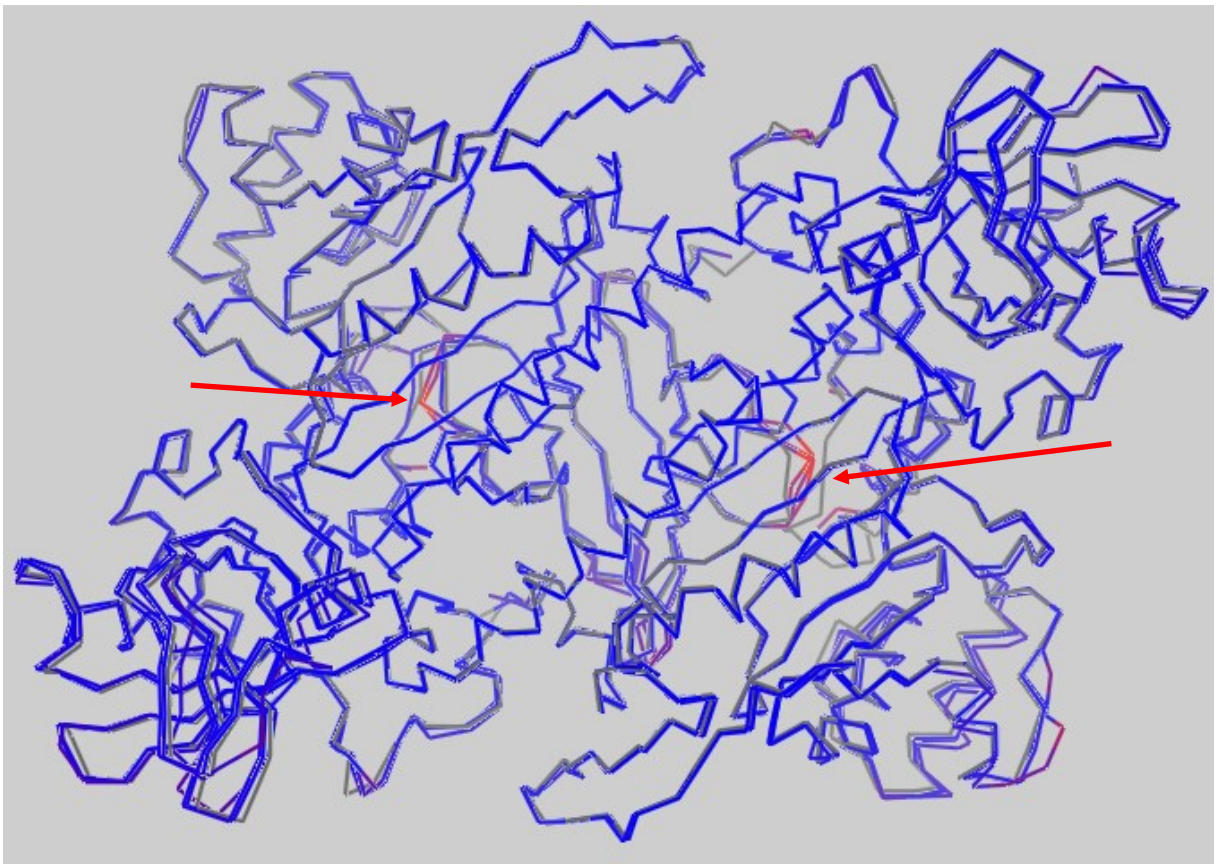

Figure 18: Combined C-alpha trace overlay coloured by RMSD.
This figure shows the C-alpha traces of four structures overlaid: the three structures from this investigation (LAM_H2, LAM_G4 and LCM_F3) and the existing Tcon LysRS structure fully-complexed with lysine (grey). The RMSD values (Å) for these alignments are 0.42, 0.28 and 0.39, respectively. A PyMOL script was used to align each of the new structures to the existing one and then colour them by local RMSD. A rainbow spectrum of colours indicates relative RMSD, starting at a dark blue for minimum RMSD ranging down to red for maximum RMSD. Red arrows indicate the loops that deviate from their conformation in the existing structure.

Now that it has been shown that there are no major deviations in the overall structural arrangement when compared to the existing Tcon LysRS model, the structures were compared to one another in the same manner. These comparisons are seen in figures 19, 20 and 21.

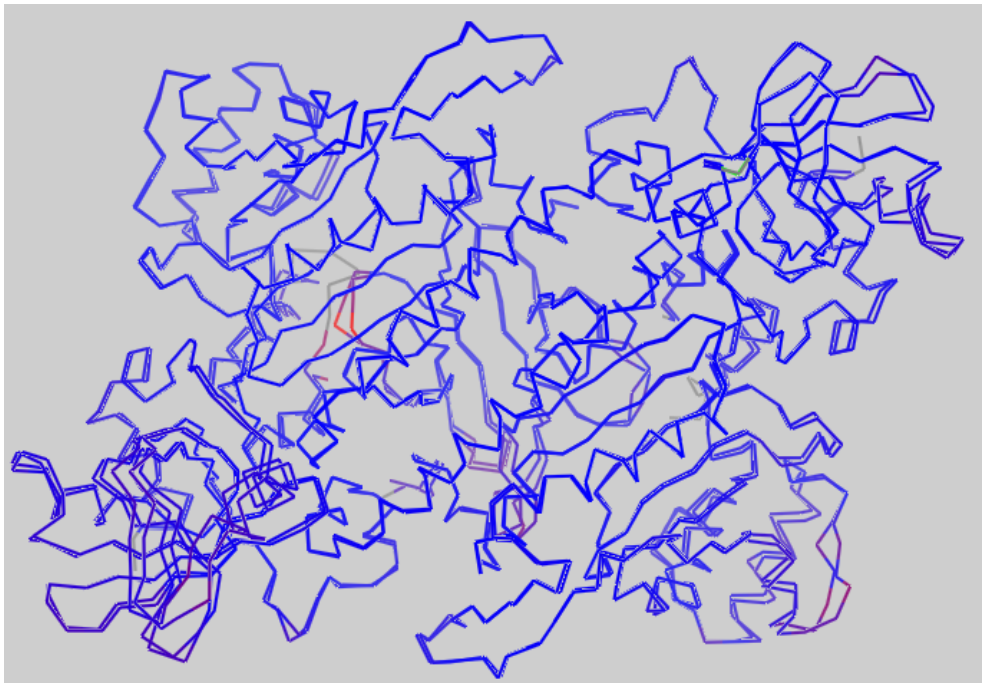

Figure 19: C-alpha trace overlay coloured by RMSD (LAM_H2 v LAM_G4).
This figure shows the C-alpha traces of LAM_H2 and LAM_G4. The overall RMSD for this alignment is 0.34 Å. A PyMOL script was used to align the two structures and then colour them by local RMSD. A rainbow spectrum of colours indicates relative RMSD, starting at a dark blue for minimum RMSD ranging down to red for maximum RMSD.

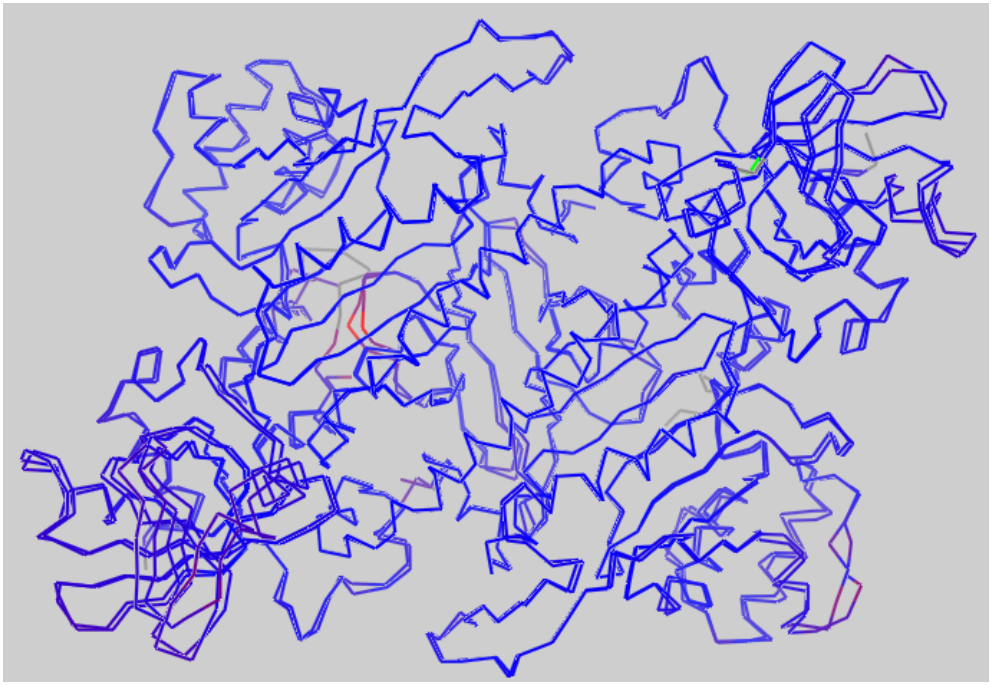

Figure 20: C-alpha trace overlay coloured by RMSD (LAM_H2 v LCM_F3).
This figure shows the C-alpha traces of LAM_H2 and LCM_F3. The overall RMSD for this alignment is 0.39 Å. A PyMOL script was used to align the two structures and then colour them by local RMSD. A rainbow spectrum of colours indicates relative RMSD, starting at a dark blue for minimum RMSD ranging down to red for maximum RMSD.

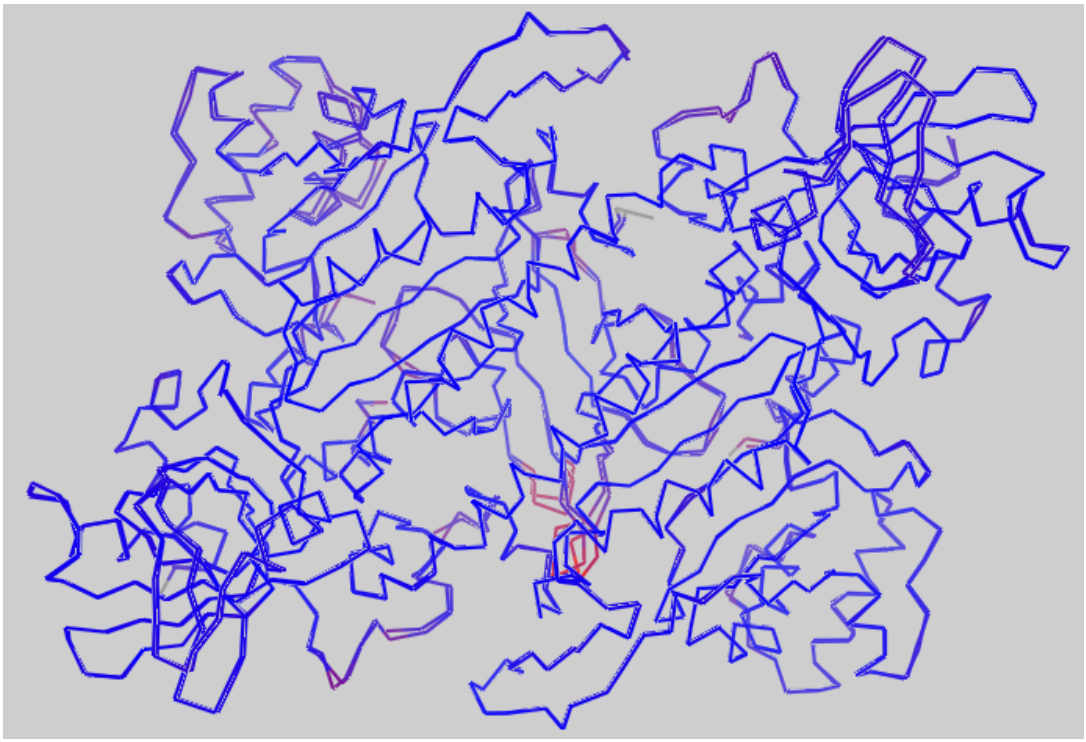

Figure 21: C-alpha trace overlay coloured by RMSD (LAM_G4 v LCM_F3). This figure shows the C-alpha traces of LAM_G4 and LCM_F3. The overall RMSD for this alignment is 0.21 Å. A PyMOL script was used to align the two structures and then colour them by local RMSD. A rainbow spectrum of colours indicates relative RMSD, starting at a dark blue for minimum RMSD ranging down to red for maximum RMSD.

As with the previous overlays, there appear to be no significant deviations between these structures – most discrepancies seen are negligible, on the external surface of the protein and/or contained to short (~1-3 residue) sections. There is one exception, aside from the loops already mentioned, which is in the comparison between LCM_F3 and LAM_G4 (Figure 21). Two adjacent areas are distinctly displaced from one another in the alignment, as highlighted in Figure 22. Due to the proximity of the first area (residues ~310-315) to the ATP-binding site, this discrepancy could potentially be significant.

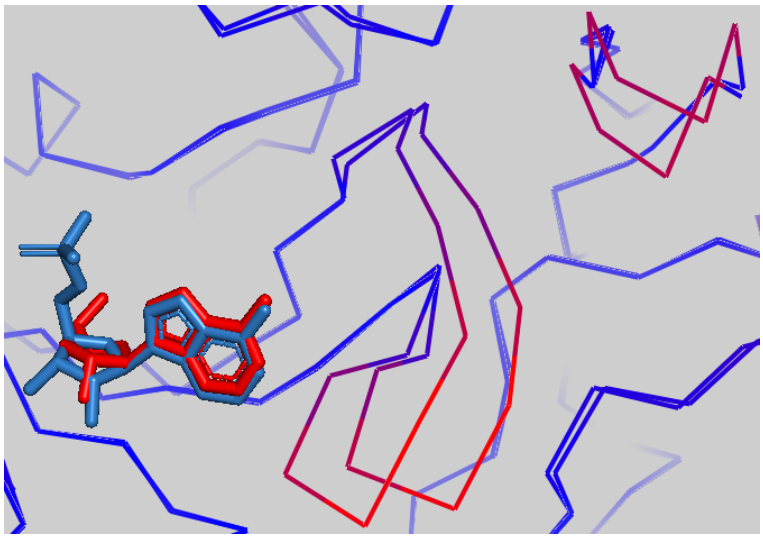

Figure 22: Area of higher deviation (LAM_G4 v LCM_F3).
This shows a zoomed-in view of the area identified in Figure 21, with the bound ligand from each structure (AMP from LAM_G4, light blue; Cordycepin from LCM_F3, red) visible to highlight the proximity of these deviations to the ATP-binding site.

As shown in figures 23 and 24, this discrepancy is seen only in chain A: LAM_G4 and the existing structure have these sections of residues (~310-315 and ~270-275) further away from the ATP-binding site than in LCM_F3 and LAM_H2. If this was observed in both chains, it could be suggested that this is part of a conformational change to promote non-optimal binding of cordycepin, a non-native ligand, and the lysyl-AMP intermediate due to a lack of tRNA with which to continue the reaction. But it is seen only in chain A, despite all of the structures except LAM_H2 containing the same bound ligand in both chains.

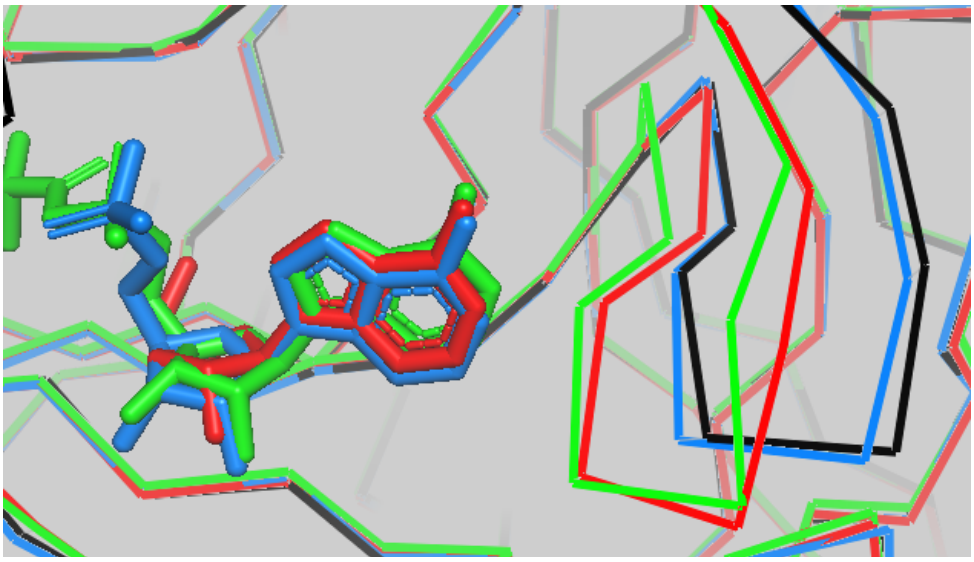

Figure 23: Combined C-alpha trace overlay (Chain A).
This figure shows the C-alpha traces of four structures overlaid: the three structures from this investigation (LAM_H2; green, LAM_G4; blue and LCM_F3; red) and the existing Tcon LysRS structure fully-complexed with lysine (black). This image is zoomed in on the area within chain A where a discrepancy has been identified (residues ~310-315).

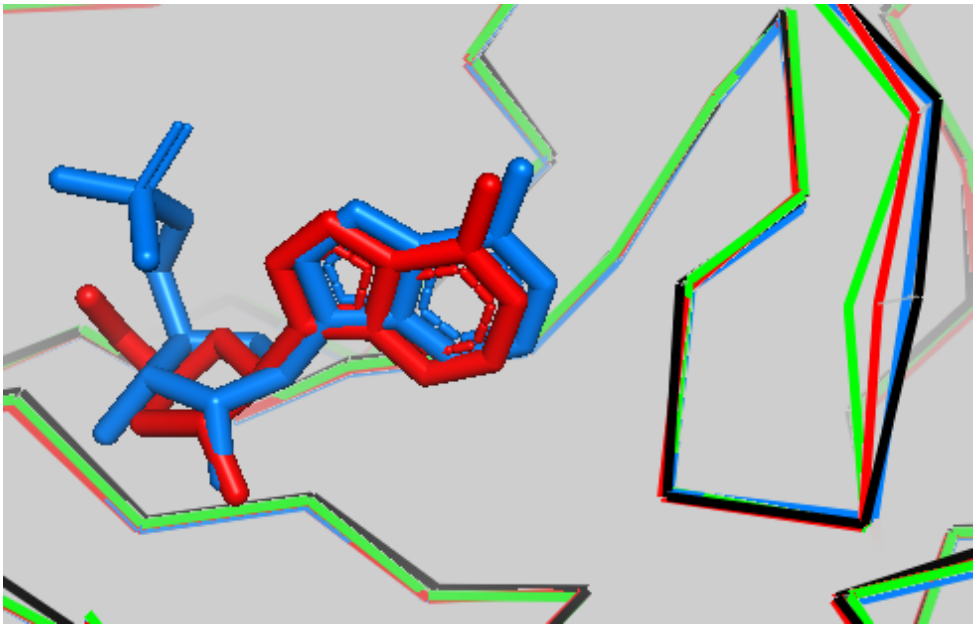

Figure 24: Combined C-alpha trace overlay (Chain B).
This figure shows the C-alpha traces of four structures overlaid: the three structures from this investigation (LAM_H2; green, LAM_G4; blue and LCM_F3; red) and the existing Tcon LysRS structure fully-complexed with lysine (black). This image is zoomed in on the area within chain B where a discrepancy is seen in the equivalent area in chain A (residues ~310-315).

Every chain was aligned together, as seen in Figure 25, and this region was then examined more closely. Out of both chains in every structure, only chain A of LAM_G4 and chain A of the lysine-only structure has this region further away from the ATP-binding site. As mentioned before, both structures contain the same ligands bound in both chains. This suggests that the alternative conformation observed in chain A of the lysine-only structure reflects the flexibility in this region when the ATP-binding site is empty. In LAM_G4, the alternative conformation in chain A therefore may indicate sub-optimal binding of the ligand in that chain.

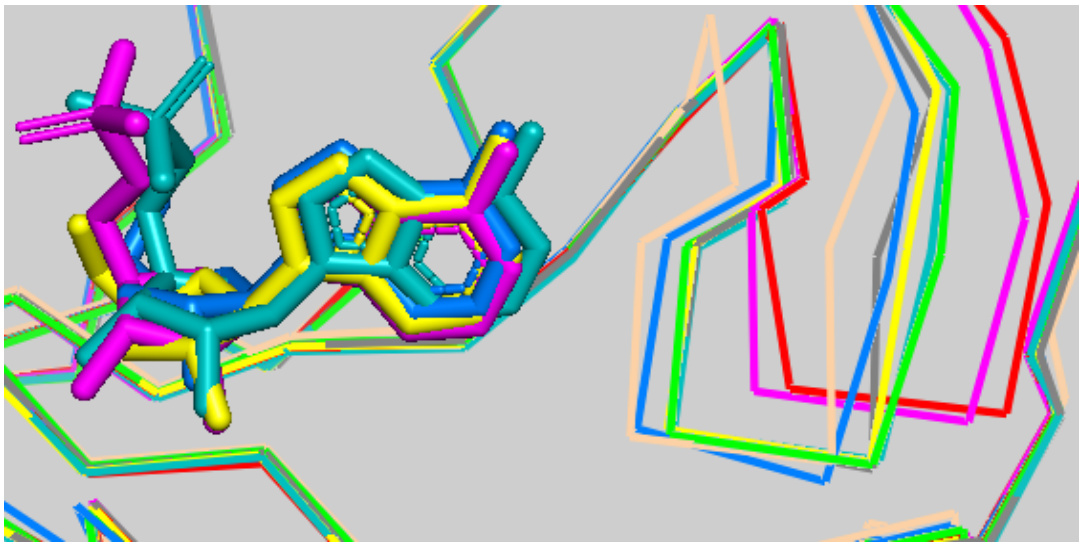

Figure 25: Combined C-alpha trace overlay (All chains).
This figure shows the C-alpha traces of all chains in four structures overlaid. The structures solved in this investigation (LCM_F3 chain A, blue; LCM_F3 chain B, yellow; LAM_G4 chain A, magneta; LAM_G4 chain B, teal; LAM_H2 chain A, wheat; LAM_H2 chain B, grey) and the existing Tcon LysRS structure fully-complexed with lysine (Chain A, red; chain B, green). This image is zoomed in on the area of the discrepancy (residues ~310-315).

Alignment	RMSD (Å)
Existing – LCM_F3	0.39
Existing – LAM_H2	0.42
Existing – LAM_G4	0.28
LCM_F3 – LAM_H2	0.39
LCM_F3 – LAM_G4	0.21
LAM_H2 – LAM_G4	0.34

Table 14: Compiled alignment RMSD values.
This figure shows RMSD values of the alignments made in this section. 'Existing' refers to the existing structure containing Tcon LysRS fully complexed with lysine.

As seen in Table 14, the most closely aligned structures are LCM_F3 (Cordycepin) and LAM_G4 (AMP). This makes sense, as cordycepin is intended to mimic ATP/AMP.

LAM_H2

In the LAM_H2 crystal structure the two molecules in the asymmetric unit are very similar but not identical: they differ primarily in two loop regions and their respective active sites as described below. A cartoon representation of the entire structure is seen below in Figure 26 with each chain coloured differently, indicating the secondary structural composition of the protein as well as the positions of the active site relative to the rest of the macromolecule. The overall fold of the structure is seen to be as previously described.

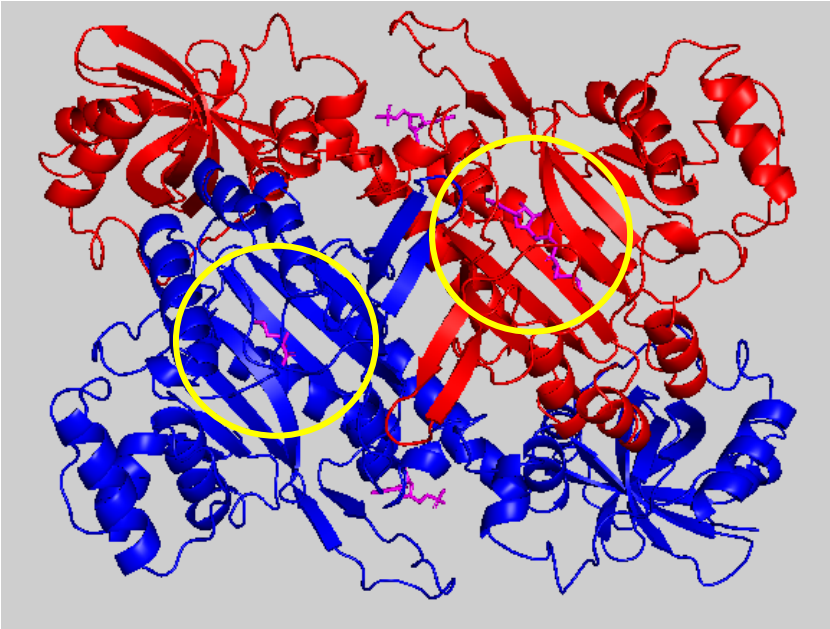

Figure 26: Overview of LAM_H2.
This image shows an overview of the LAM_H2 structure, produced in PyMOL. Chain A is in red, chain B is in blue and the ligands are magenta. The active sites are indicated by yellow rings.

Figure 27 displays another C-alpha trace comparison, this time between the two monomers in LAM_H2, in the same orientation as the previous traces.

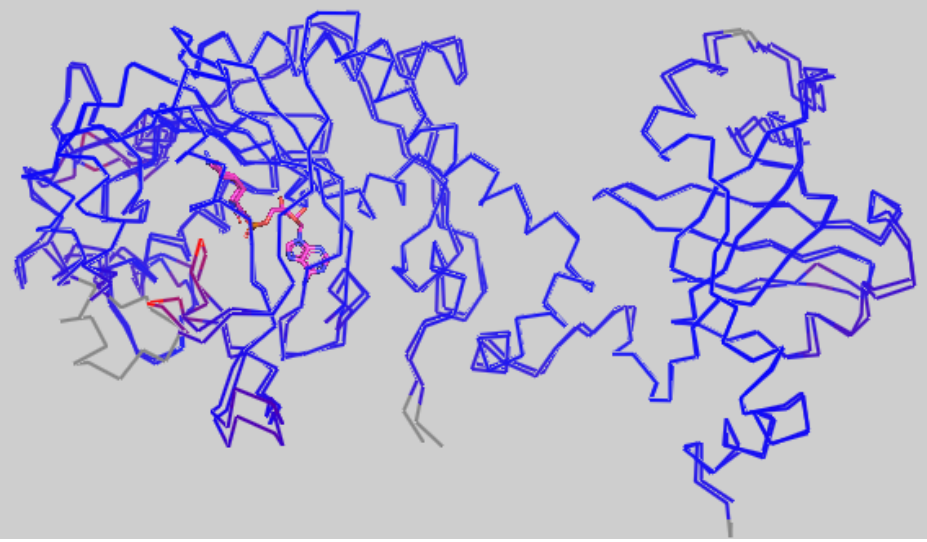

Figure 27: C-alpha trace overlay coloured by RMSD (LAM_H2_A v LAM_H2_B).
This figure shows the C-alpha traces of chain A and chain B of LAM_H2. The RMSD for this alignment is 0.53 Å. A PyMOL script was used to align the two structures and then colour them by local RMSD. A rainbow spectrum of colours indicates relative RMSD, starting at a dark blue for minimum RMSD ranging down to red for maximum RMSD. Grey indicates areas that were excluded from the alignment. The lysyl-AMP intermediate is visible to show the location of the substrate binding sites.

A significant difference between the two copies of LysRS1 in the LAM_H2 crystal structure is the active site content. In chain A the active site contains electron density consistent with a bound lysyl-AMP intermediate (Figure 28), whereas in chain B lysine alone appears to be bound (Figure 29). This result was unexpected as the crystallisation content contained both lysine and ATP substrates at concentrations more than sufficient to fully occupy both active sites.

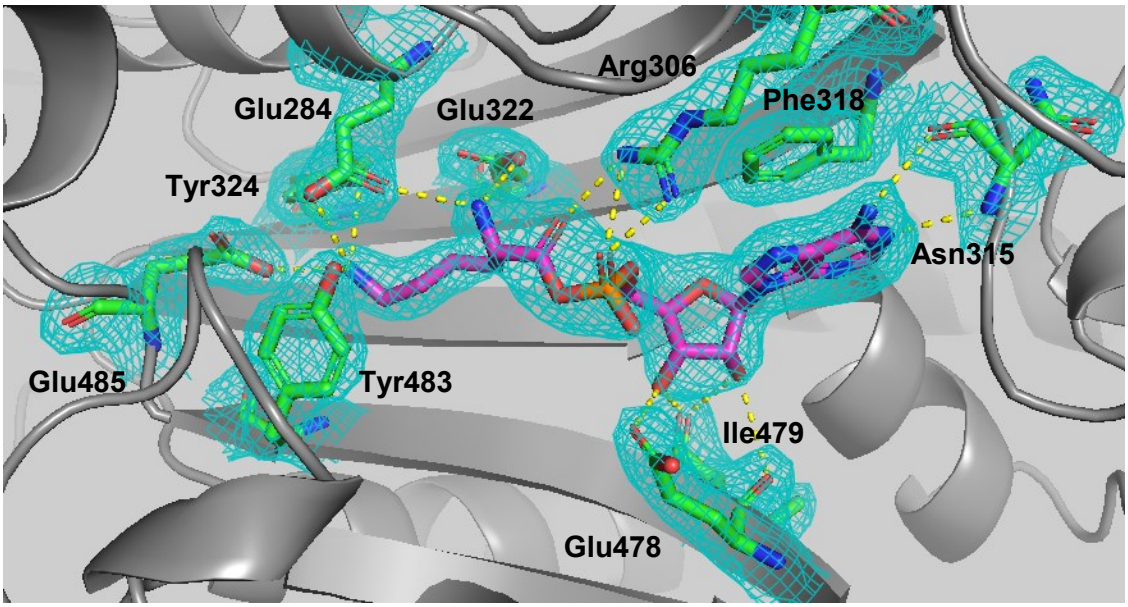

Figure 28: LAM_H2, active site of chain A.
This image shows the active site of chain A in the LAM_H2 structure. The lysyl-AMP intermediate is highlighted in purple, whereas binding residues are shown green and labelled in black. Hydrogen bonds are shown as yellow dotted lines.

The active site electron density in molecule A (Figure 28) was very clear and is consistent with the presence of a lysyl-AMP intermediate interacting with mainly side chains of a number of residues in the active site of chain A. Tyr324, Tyr483, Glu485 and Glu284 all form hydrogen bonds with the epsilon amino group from the 'lysyl' moiety of the intermediate. Glu284 forms a second hydrogen bond, together with Glu322, with the alpha amino group of that area. Arg306 hydrogen bonds with both the carboxyl group and the phosphate itself. On the ribose, Glu478 and Ile479 (main chain carbonyl) again can be seen to both hydrogen bond with both the 2'- and 3'-hydroxyl groups. The Asn315 main chain amide group hydrogen bonds with the adenine ring and the adenine amino group.

As might be expected, the binding interactions for the lysyl-AMP intermediate essentially consist of the sum of the interactions seen later in LAM_G4 for lysine and AMP separately.

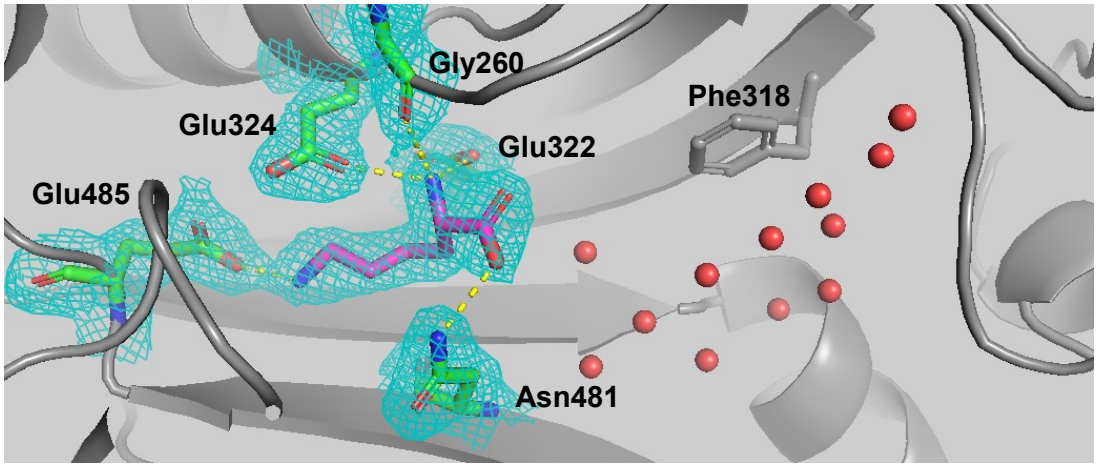

Figure 29: LAM_H2, active site of chain B.
This image shows the active site of chain B in the LAM_H2 structure. The lysine ligand is highlighted in purple, whereas binding residues are shown green and labelled in black. Hydrogen bonds are shown as yellow dotted lines. Phe318 is shown as a reference point. Density interpreted as bound water molecules is shown as red spheres.

The active site of chain B (Figure 29) exhibits electron density consistent only with a bound lysine and the other half of the binding pocket appearing to be occupied by waters as seen in the previously reported 'lysine only' LysRS structures (McCready-Fallon, 2018). This conclusion was reached because the density data in that pocket is disconnected, failed to improve when any of the potential ligands (ATP, AMP, intermediate) were added in, and nor does it appear to be in the expected orientation for any of those potential ligands.

The lysine substrate is bound via a cluster of hydrogen bonds between its alpha amino group and the nearby Glu284, Gly260 (main chain carbonyl) and Glu322 residues. This Gly260 is part of a loop that seems to appear in either of two conformations and so this hydrogen bond, among the structures produced in this investigation which otherwise have this loop in the alternate

conformation, is only seen in chain B of LAM_H2. It is possible that this loop works to anchor the lysine in place prior to the binding of the ATP, at which point the reaction occurs and the loop moves away from the active site into its other conformation. One hypothesis is that this could potentially promote the binding of tRNA following ATP hydrolysis. Finally, the bound lysine carboxyl group also forms a hydrogen bond with the side chain amide nitrogen of Asn481.

In addition to hydrogen bonding, the site has a number of residues that are involved in hydrophobic or Van Der Waals (VDW) contacts with the bound ligand(s). These residues contribute to shape complementarity of the enzymatic site to the substrates and, crucially, help prevent incorrect molecules from binding in place of the desired substrate. They can be seen below in figures 30, 31 and 32.

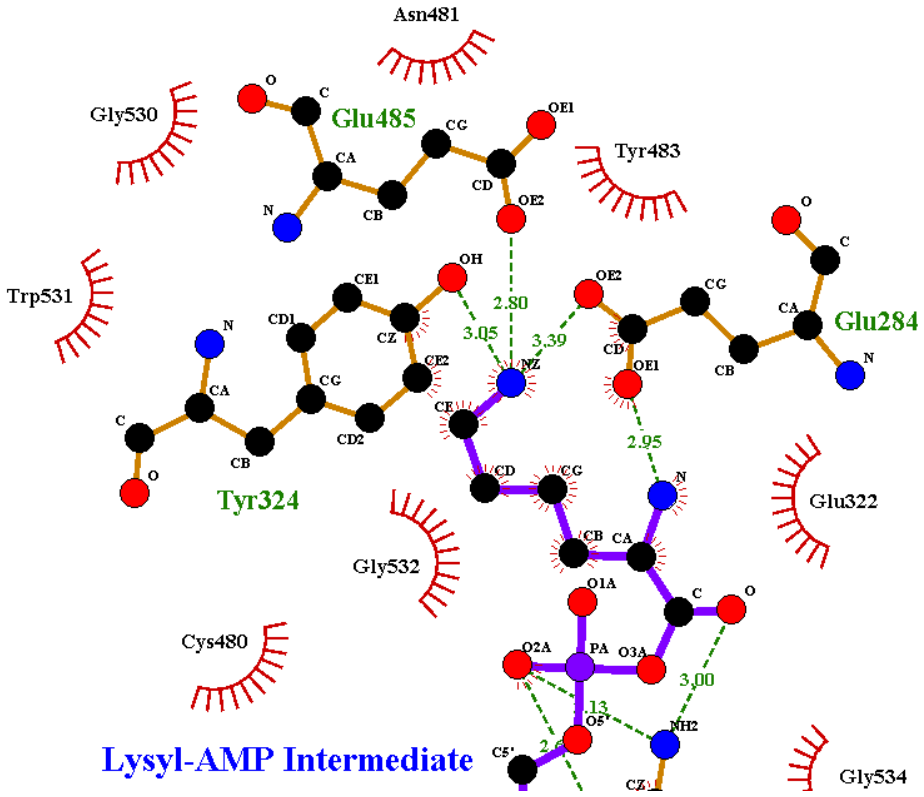

Figure 30: LAM_H2 chain A lysine-binding site in LigPlot.
This is a schematic of the chain A lysine-binding site of LAM_H2, prepared using LigPlot. It shows the hydrogen bonds (green dotted lines) more clearly and provides information on residues/atoms involved in either hydrophobic or VDW contacts (red crescents) with the bound ligand.

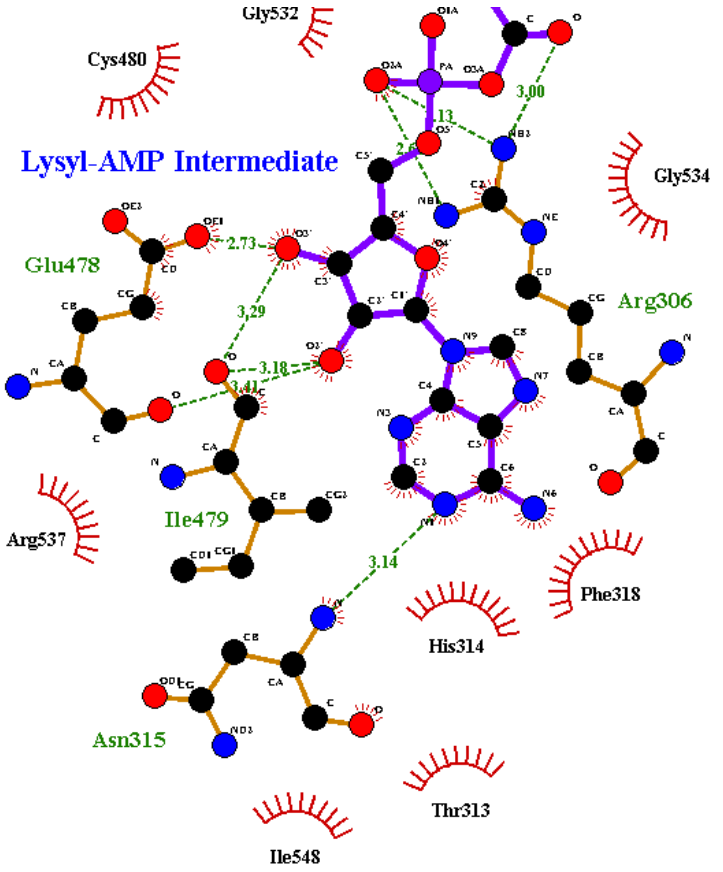

Figure 31: LAM_H2 chain A ATP-binding site in LigPlot.
This is a schematic of the chain A ATP-binding site of LAM_H2, prepared using LigPlot. It shows the hydrogen bonds (green dotted lines) more clearly and provides information on residues/atoms involved in either hydrophobic or VDW contacts (red crescents) with the bound ligand.

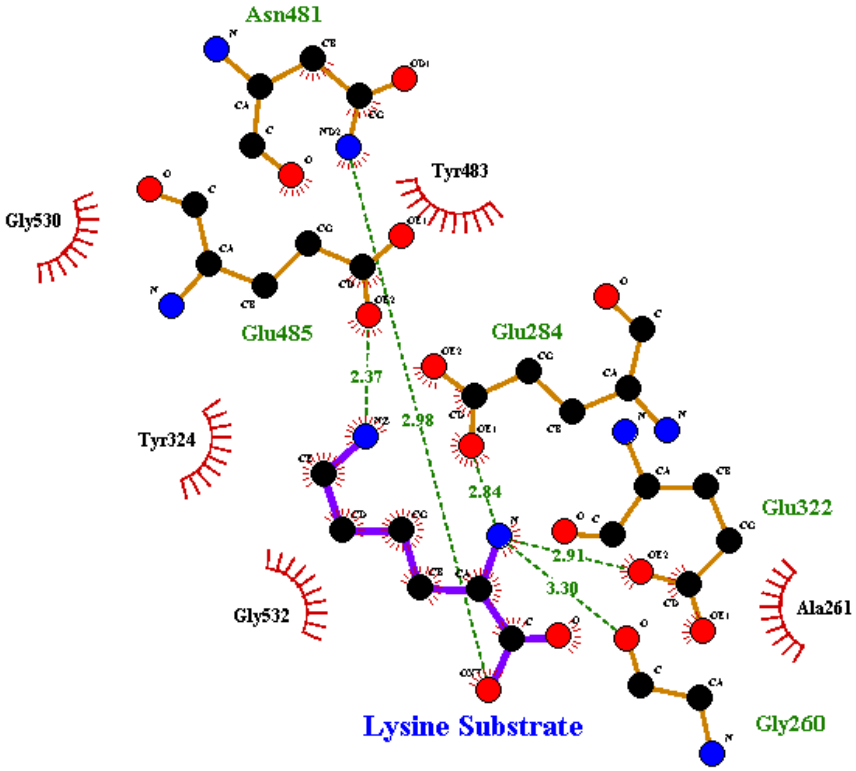

Figure 32: LAM_H2 chain B lysine-binding site in LigPlot.
This is a schematic of the chain B lysine-binding site of LAM_H2, prepared using LigPlot. It shows the hydrogen bonds (green dotted lines) more clearly and provides information on residues/atoms involved in either hydrophobic or VDW contacts (red crescents) with the bound ligand.

Table 15 has been prepared to summarise the residues in each of these sites involved in hydrophobic or VDW contacts with the ligand.

Site	Bound ligand	Hydrophobic or VDW contacts
Chain A lysine-binding	Lysyl-AMP Intermediate	Glu284, Glu322, Tyr324, Asn481, Tyr483, Gly530, Trp531, Gly532
Chain A ATP-binding	Lysyl-AMP Intermediate	Arg306, Thr313, His314, Asn315, Phe318, Glu478, Ile479, Cys480, Gly534, Arg537, Ile548
Chain B lysine-binding	Lysine	Ala261, Glu284, Glu322, Tyr324, Asn481, Tyr483, Glu485, Gly530, Gly532
Chain B ATP-binding	N/A (Water)	N/A

Table 15: Summary of hydrophobic or VDW contacts in LAM_H2 catalytic sites.
This table lists the residues that are involved in hydrophobic or VDW contacts with the bound substrates in each catalytic site as reported by LigPlot.

Structural changes in and around the active site

Comparative analysis of the LysRS1 active sites has highlighted two loop regions of significant interest in these structures. The first of these, consisting of residues 258-265 (AGGAAARP), will be termed ‘Loop-Alpha’, and the second which is an external surface loop roughly consisting of residues 500-514 (LRDREKGDDEAMSID), will be referred to as ‘Loop-Beta’.

Loop-Alpha is inherently interesting because it contains a residue (Gly260) that is ligand-binding in LAM_H2 chain B and the previous structures of this enzyme complexed with lysine, yet far apart from the active site in the other structures solved in this investigation. This is consistent with a conformational change during enzyme activity. Loop-Beta is often poorly defined in the existing crystal structures of this enzyme, with no corresponding density in most chains and only patchy density where there is any. While this is not unusual for an external loop that is likely very flexible, Loop-Beta is important because its location is adjacent to the position of Loop-Alpha; that is, their movements are potentially linked as part of the same concerted conformational change.

In the structures solved during this investigation there are two apparent, distinct conformations for Loop-Alpha which will simply be referred to as ‘1’ and ‘2’. Conformation 1 places Gly260 close to the bound lysine, where it forms a main chain hydrogen bond to the alpha-amino group. This is shown below in Figure 33. When this conformation is observed, Loop-Beta does have representation in the density – whether patchy or clear, there is an indication of where Loop-Beta resides in structures where Loop-Alpha is in conformation 1.

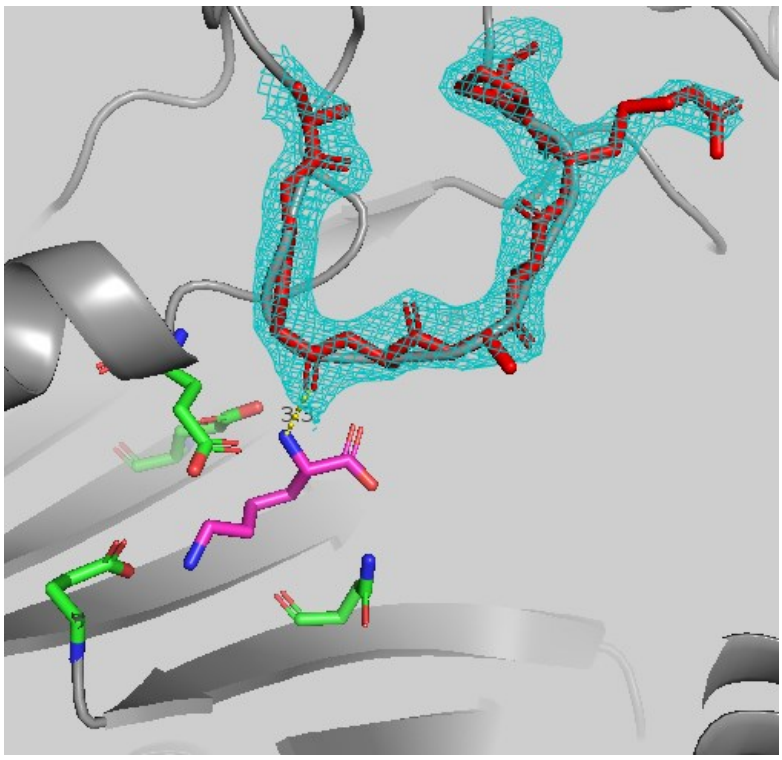

Figure 33: LAM_H2, Loop-Alpha of chain B – "Conformation 1".
This image shows the position of Loop-Alpha, highlighted in red, with respect to the active site of chain B. A dotted line shows the distance in Å between Gly260 and the alpha-amino group of the lysyl-AMP intermediate.

Conformation 2 places Gly260 further from the active site and, when this conformation of Loop-Alpha is observed, there is a complete absence of any density that could conceivably represent Loop-Beta. When Loop-Alpha is in conformation 1, Loop-Beta occupies the space that Loop-Alpha would occupy were it in conformation 2. Conformation 2 is shown in Figure 34 below.

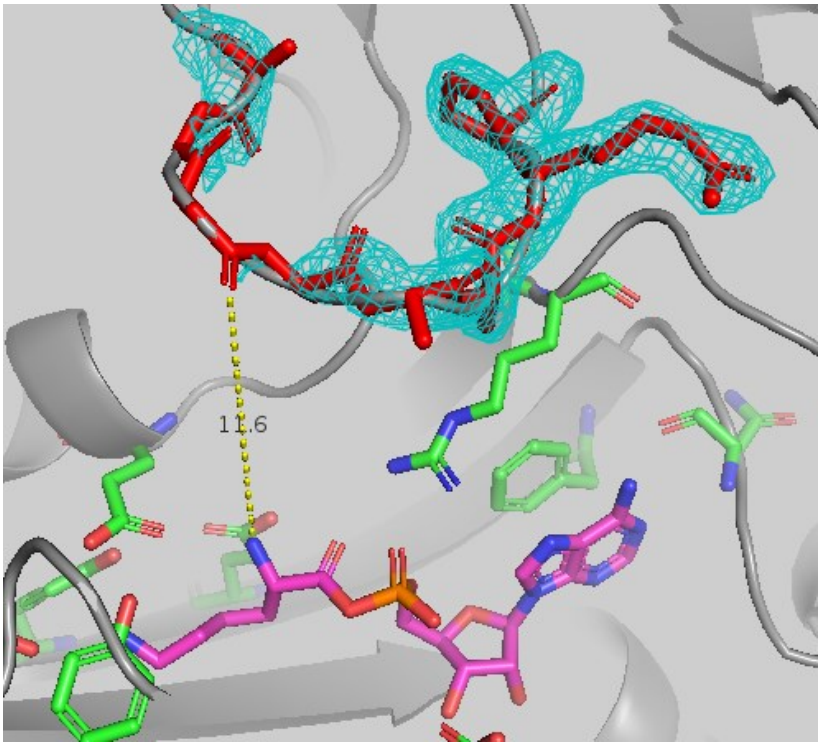

Figure 34: LAM_H2, Loop-Alpha of chain A – "Conformation 2".
This image shows the position of Loop-Alpha, highlighted in red, with respect to the active site of chain A. A dotted line shows the distance in Å between Gly260 and the alpha-amino group of the lysyl-AMP intermediate.

In LAM_H2, Loop-Alpha is in conformation 1 in chain B where only lysine is bound in the active site. This is in concurrence with the previously reported structures of Tcon LysRS1, where Loop-Alpha is also seen in conformation 1 in chains where lysine alone is present.

In contrast, in chain A where the intermediate is present Loop-Alpha can be seen clearly to be in conformation 2 despite slightly patchier density for it. The 'semi-complexed' Tcon LysRS1 (McCready-Fallon, 2018), in the chain where there are no ligands bound in the active site, has Loop-Alpha in a position resembling this conformation 2. This may indicate that the binding of lysine induces a conformational change that promotes the binding of ATP. After ATP binds and undergoes hydrolysis to AMP and then condensation to lysyl-AMP, another conformational change takes place which reverts the loops back to their original positions. This could be to promote tRNA binding or simply to, in this case where there is no tRNA to facilitate the next stage of the reaction, instead facilitate the replacement of the intermediate by another ATP – to initiate the catalytic cycle once again. Although this analysis is conjecture at this stage, the role of these loops will be returned to in each of the analyses of structures generated in this investigation.

Comparison with 6CHD

No structures of bovine LysRS exist in the PDB, and the sequence for this enzyme on Uniprot – F1MMK8 – is only partially validated. Even so, a BLASTp search was conducted using the F1MMK8 sequence and this returned a high similarity (Table 16) with mitochondrial human LysRS (Q15046-2) for which there are two structures (3BJU and 6CHD) in the PDB. 3BJU has already been mentioned earlier in this investigation as a tetrameric human LysRS enzyme with ATP and lysine bound to each chain, and 6CHD is a recently-uploaded structure to the PDB but for which there is no associated publication. In this structure human LysRS has an intermediate analogue (the phosphorus atom is instead a sulphur) bound in both active sites of the dimer. Due to its high sequence similarity to bovine LysRS and with no obvious alternative, 6CHD was used as a model in comparisons of LAM_H2 with a mammalian intermediate-containing LysRS structure. The aim was to identify any significant differences that could prove key to the development of a drug that distinguishes between parasite and host.

```
#=======================================
#
# Aligned_sequences: 2
# 1: 6CHD
# 2: F1MMK8
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 578
# Identity:     545/578 (94.3%)
# Similarity:   562/578 (97.2%)
# Gaps:           2/578 ( 0.3%)
# Score: 2876.5
#
#
#=======================================
```

Table 16: Summary of alignment between 6CHD and F1MMK8.

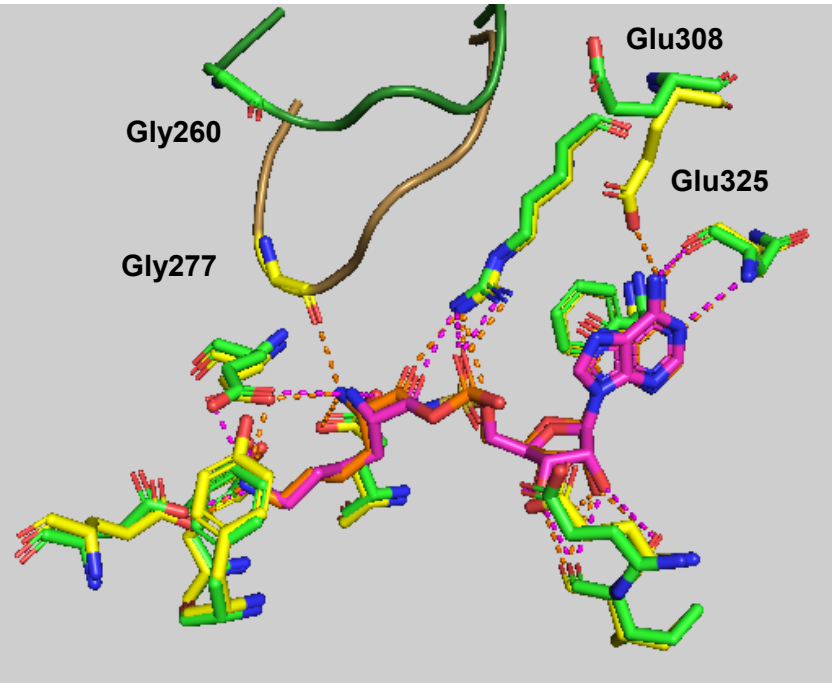

Figure 35: LAM_H2 (green & purple) v 6CHD (yellow & orange).
This image shows the intermediate-occupied active site of LAM_H2 overlaid with the active site of 6CHD (Human LysRS). LAM_H2 is coloured with green carbons, and the intermediate is magenta, whereas 6CHD is coloured with yellow carbons and the intermediate analogue is orange. Hydrogen bonds match the colour of the ligand participating in them. Loop-Alpha is shown in cartoon form.

As can be seen in Figure 35, the intermediate analogue in 6CHD fits into the active site in a very similar manner to the actual intermediate observed in LAM_H2. Most residues of interest are the same but for two notable exceptions – Glu308 (LAM_H2) / Glu325 (6CHD) and Gly260 (LAM_H2) / Gly277 (6CHD). In LAM_H2, the glutamate at position 308 is twisted away from the active site and so can be surmised to not be involved in ligand binding. This is the case in every *T. congolense* LysRS structure in this investigation and the two previously reported which only contained lysine.

The second major difference between these two active sites is the glycine, part of 'Loop-Alpha'. As seen in Figure 35, Loop-Alpha is in conformation 1 in 6CHD. In LAM_H2, as mentioned earlier, it is in conformation 1 when lysine is alone in the active site and in conformation 2 when the intermediate is bound. A comparison showing the positions of this loop in each structure is seen in Figure 36.

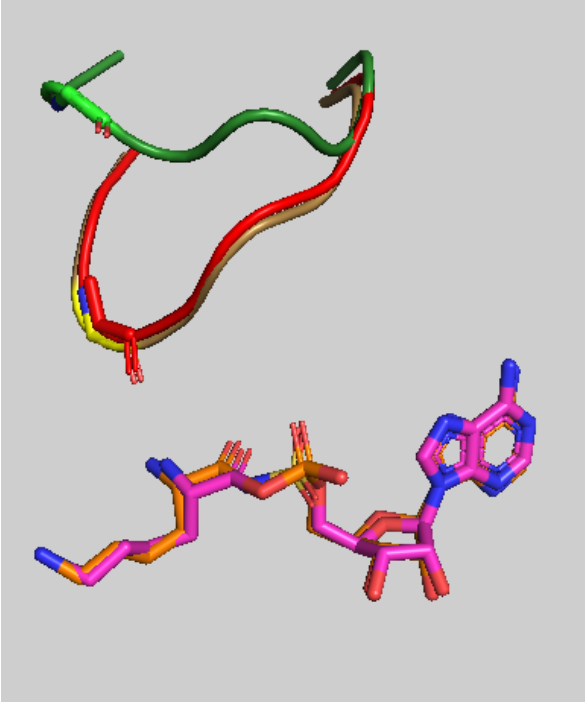

Figure 36: LAM_H2 v 6CHD Loop-Alpha comparison.
This image is a comparison between LAM_H2 and 6CHD that shows the two different conformations of Loop-Alpha. LAM_H2 chain A (intermediate-bound) is green, chain B (lysine-bound) is red, and 6CHD (intermediate analogue) is yellow. The intermediate and analogue are included as a point of reference to the previous figure.

It is not immediately clear why Loop-Alpha is in different conformations in these structures. One possibility is simply that this is a major difference between the mammalian and trypanosome enzymes. It could also be that the enzyme is distinguishing between the (presumed) real lysyl-AMP intermediate in LAM_H2, which would facilitate normal enzyme function, and the sulphur-containing intermediate analogue used in 6CHD that inhibits the enzyme.

Free AMP

The LAM_H2 structure contains a ‘free’ AMP bound to each of the active sites, as illustrated in Figures 37 and 38. In both instances the AMP adenyl ring pi-stacks with Tyr349 and forms hydrogen bonds with Lys355, although the ribose 2’-oxygen from free AMP in chain A also forms a hydrogen bond with Tyr349. This will be discussed in more detail in the following section on LAM_G4.

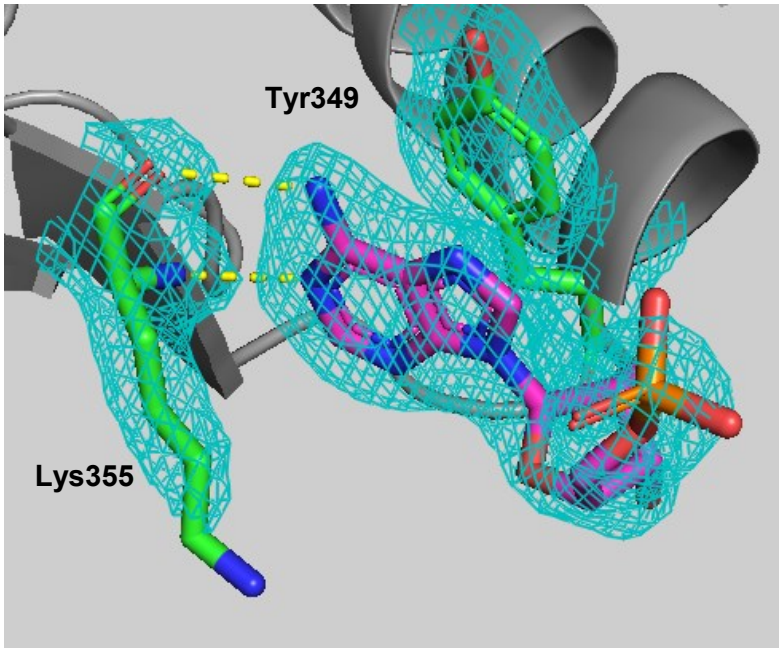

Figure 37: LAM_H2, free AMP at chain A.
The free AMP is shown in magenta, while the two residues of interest are in green and labelled in black.

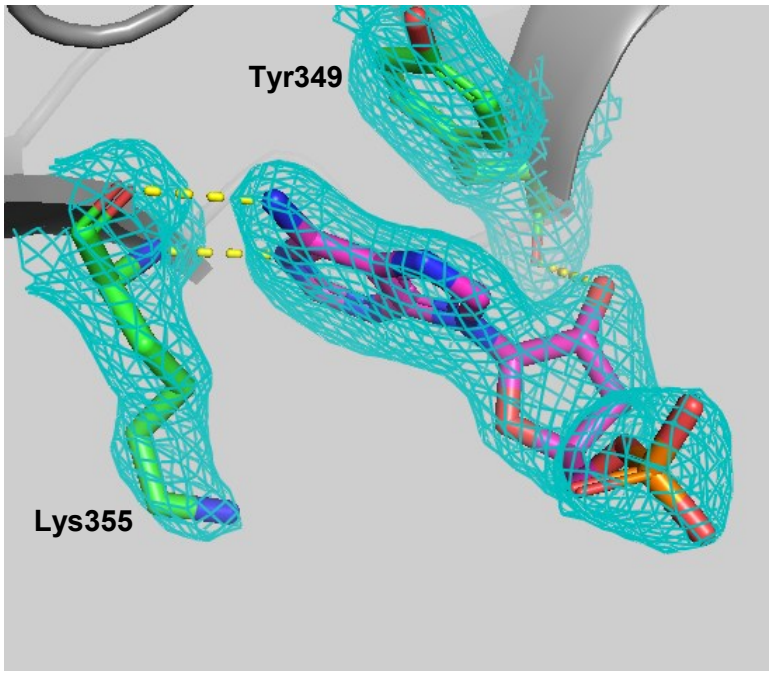

Figure 38: LAM_H2, free AMP at chain B.
The free AMP is shown in magenta, while the two residues of interest are in green and labelled in black. The image is at a slightly different angle than the previous Figure in order to show the extra hydrogen bond between free AMP and the Tyr349.

LAM_G4

The LAM_G4 crystal structure contains homodimeric LysRS1 with each active site occupied by both lysine and AMP. Figure 39 shows a cartoon rendering of the entire structure, featuring the same overall fold that was previously described, coloured by chain as before.

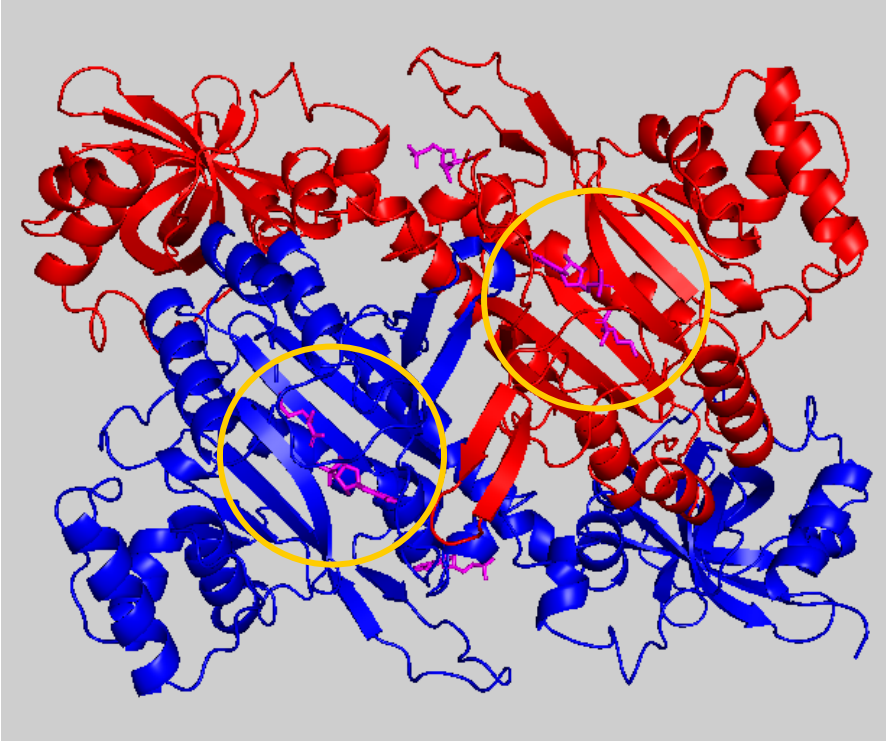

Figure 39: Overview of LAM_G4.
This image shows an overview of the LAM_G4 structure, produced in PyMOL. Chain A is in red, chain B is in blue and the ligands are magenta. The active sites are indicated by yellow rings.

In order to confirm the similarities and assess any major differences between the folding of the molecules, their C-alpha traces have been overlaid in Figure 40 below.

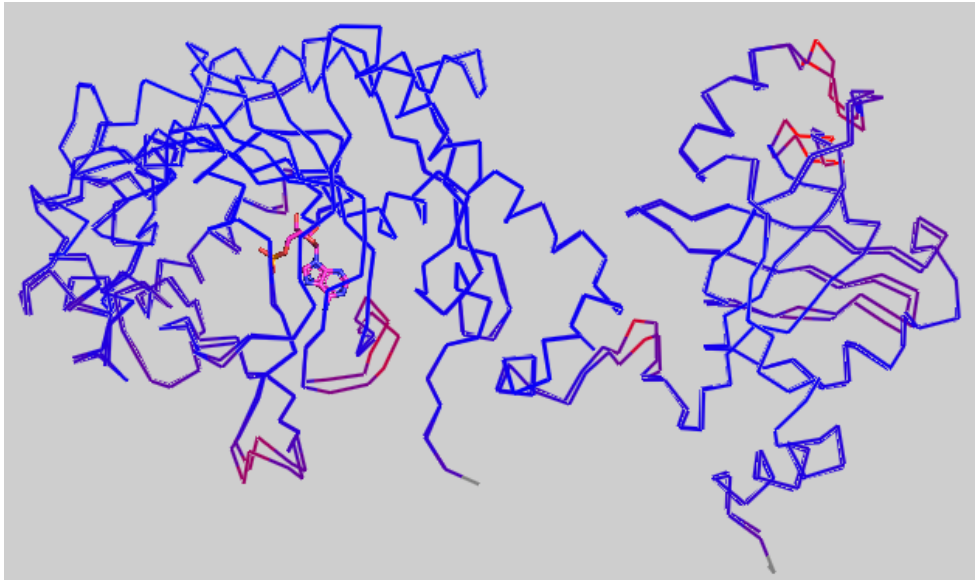

Figure 40: C-alpha trace overlay coloured by RMSD (LAM_G4_A v LAM_G4_B). This figure shows the C-alpha traces of chain A and chain B of LAM_G4. The overall RMSD of this alignment is 0.30 Å. A PyMOL script was used to align the two and then colour them by local RMSD. A rainbow spectrum of colours indicates relative RMSD, starting at a dark blue for minimum RMSD ranging down to red for maximum RMSD. Grey indicates areas that were excluded from the alignment. Bound AMP is visible to show the location of the catalytic site.

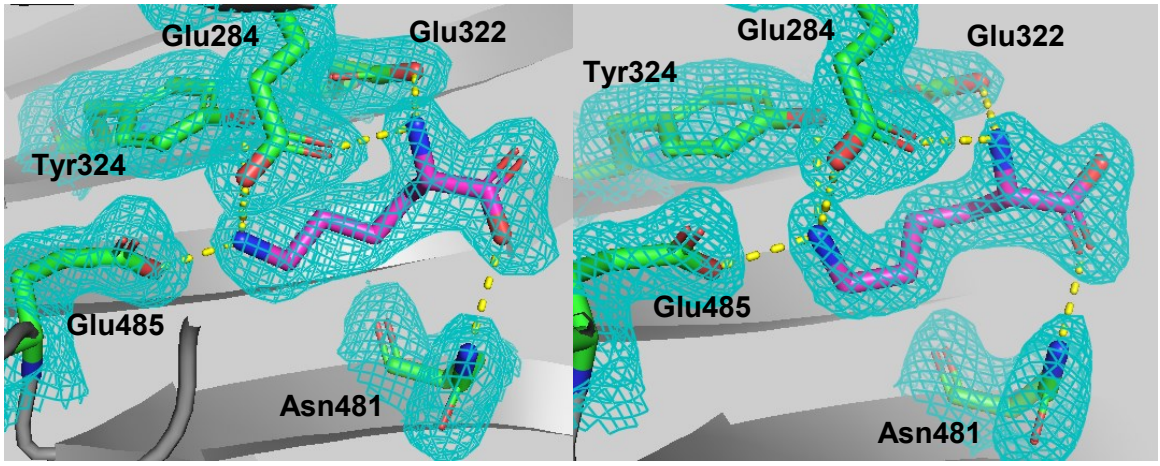

Figure 41: Lysine bound to chain A and chain B in LAM_G4. This image shows the lysine ligands bound to the active site in chain A (left) and the active site in chain B (right), in LAM_G4. Residues of interest are coloured green whilst ligands are coloured magenta. Hydrogen bonds are indicated by yellow dotted lines.

As seen in Figure 41, there is clear density for a lysine substrate in both active sites. In both cases the lysine is held by the same five residues within the active site, as also reported in a previous study (McCready-Fallon, 2018). Residues Glu284, Tyr324 and Glu485 hydrogen bond to the terminal amine group on the side chain of the lysine ligand. Glu284 also forms hydrogen bonds with the main chain alpha-amino group along with Glu322, and Asn481 hydrogen bonds the carbonyl oxygen. Information on the residues/atoms forming hydrophobic or VDW contacts with

the bound lysine ligands in these sites is provided below in figures 42 and 43, and later listed along with those in the ATP-binding site.

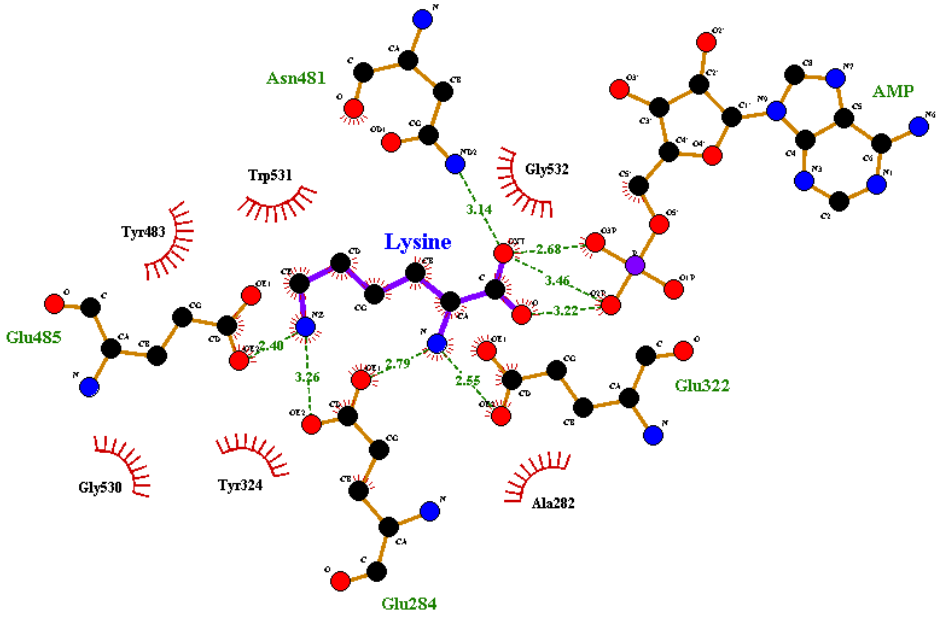

Figure 42: LAM_G4, chain A lysine-binding site.
This is a schematic of the lysine-binding site of LAM_G4 chain A, prepared using LigPlot. It shows the hydrogen bonds (green dotted lines) more clearly and provides information on residues/atoms involved in hydrophobic or VDW contacts (red crescents) with the ligand. As the Tyr324 hydrogen bond reported by PyMOL is not predicted by LigPlot, it is likely weak or has geometry that is not ideal.

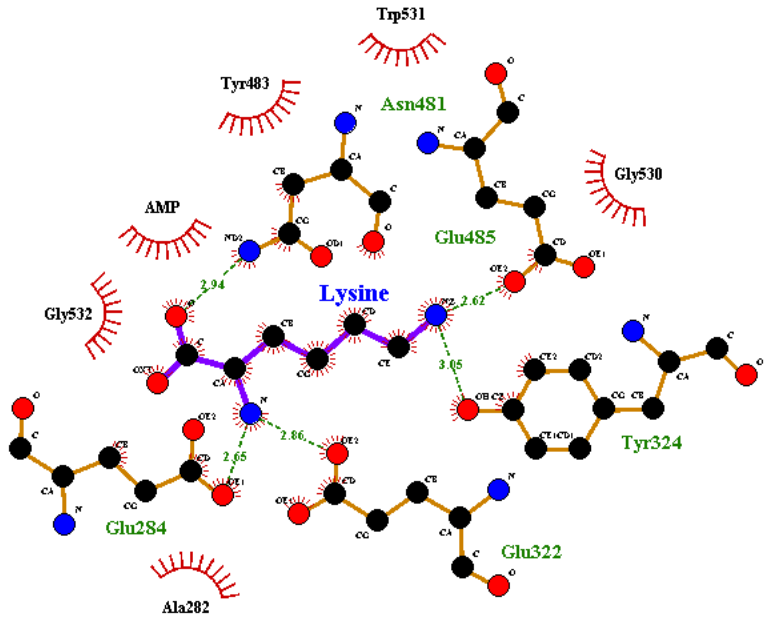

Figure 43: LAM_G4, chain B lysine-binding site.
This is a schematic of the lysine-binding site of LAM_G4 chain B, prepared using LigPlot. It shows the hydrogen bonds (green dotted lines) more clearly and provides information on residues/atoms involved in hydrophobic or VDW contacts (red crescents) with the ligand.

Unfortunately, despite the structure LAM_G4 fielding the highest resolution of the three structures, the other half of the active site has density that remains unclear mainly due to the absence of distinctive electron density for any potential phosphate groups. This made it difficult to determine whether it is AMP or ATP, or even the intermediate, which is bound in this site.

The adenine nucleobase has well-defined density in both chains and is certainly present, pi-stacking with the R-group of Phe318 as seen in 3BJU. In addition, the density supports the presence of the ribose and now indicates clearly its conformation. The phosphate region on the other hand was ambiguous – in both chains the density was insufficiently clear to define how many phosphate groups were present, and whether there is an ester linkage between the lysine ligand and this nucleotide.

During the refinement process, multiple different molecules were inserted into the site to determine the most likely ligands. Various options tested are shown in Figure 44 below.

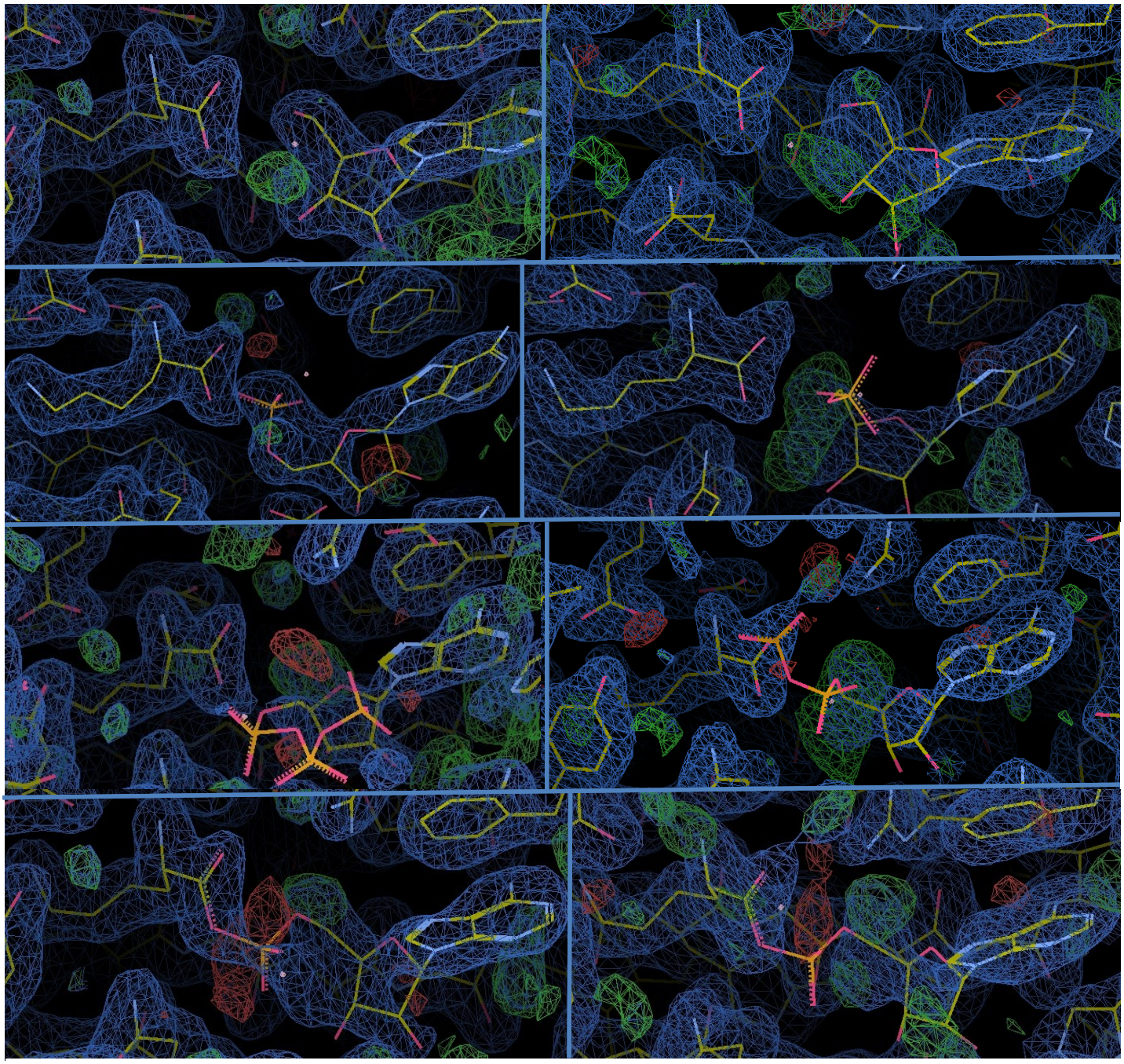

Figure 44: Investigation of potential nucleotide-based molecule.
The left column shows the active site of chain A and the right column shows the active site of chain B. The first row shows the structure with L-lysine and adenosine placed in the binding site. The second row shows the structure with L-lysine and AMP in the site. The third row shows the structure with L-lysine and ATP in the site. The final row shows the structure with the lysyl-AMP intermediate located in the site.

One other complication – or clue – was the presence of 'free' AMP molecules, one in each chain, pi-stacking with Tyr349. These are very clearly represented in the electron density and indicate that the enzyme is functioning – only ATP was added to the solution and so any AMP is most likely to have been formed as a result of enzyme activity.

However, if the enzyme is functioning, it should only be capable of reaction step 1 because there is no tRNA present to facilitate step 2. Step 2 is where AMP is released and so, if the enzyme can perform step 1 in these conditions, there should only be ATP and/or lysyl-AMP in the structure – as reactant and product, respectively, of that first step. The intermediate is known to be unstable in solution which usually keeps it from dissociating from the enzyme before the second stage begins (Onesti, 2000). However, as the rate of reaction is presumably slow at these low temperatures and there is no tRNA to permit the next step of the reaction, one potential explanation is that the intermediate is being outcompeted by the high concentration of substrates. Upon being displaced from the active site, the intermediate breaks down, hence the free AMP molecules.

In figures 45 and 46, the ribose oxygen of the free AMP in chain B can be seen to form a hydrogen bond with the main chain carbonyl of Tyr349, but otherwise those in both chain A and B form two hydrogen bonds with Lys355 and pi-stacking interactions with that Tyr349. Figures 47 and 48 detail the residues that form hydrophobic or VDW contacts with the free AMP in each chain. These ligands are on the external surface of the protein and do not bind as strongly as those in the catalytic site.

As seen earlier in the overall structure image in Figure 39, the catalytic site is relatively close by and so this ‘free AMP’ site could theoretically play a role in enzyme function.

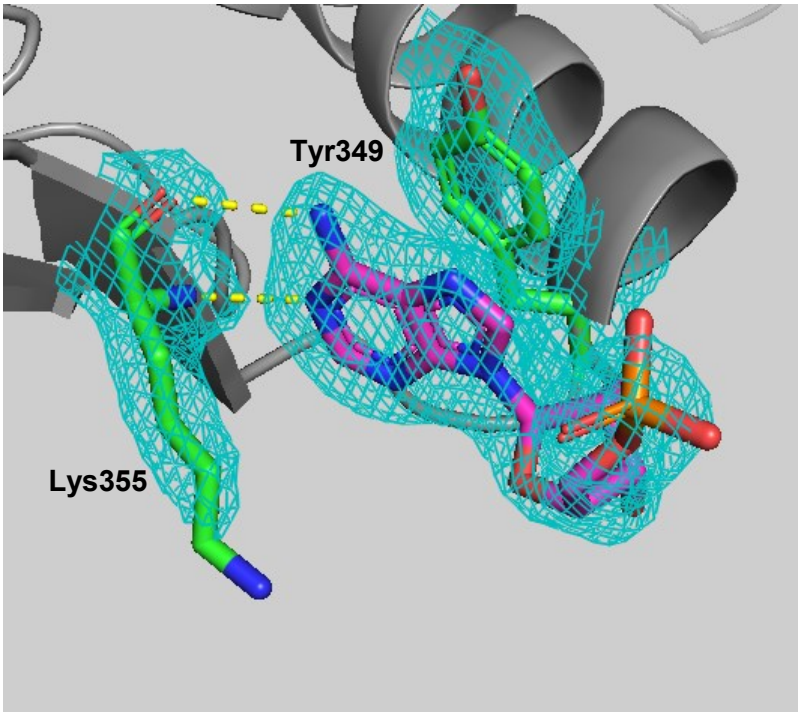

Figure 45: Free AMP in chain A.
In chain A, the ‘free’ AMP ligand (magenta) forms two hydrogen bonds with Lys355 and pi-stacks with Tyr349.

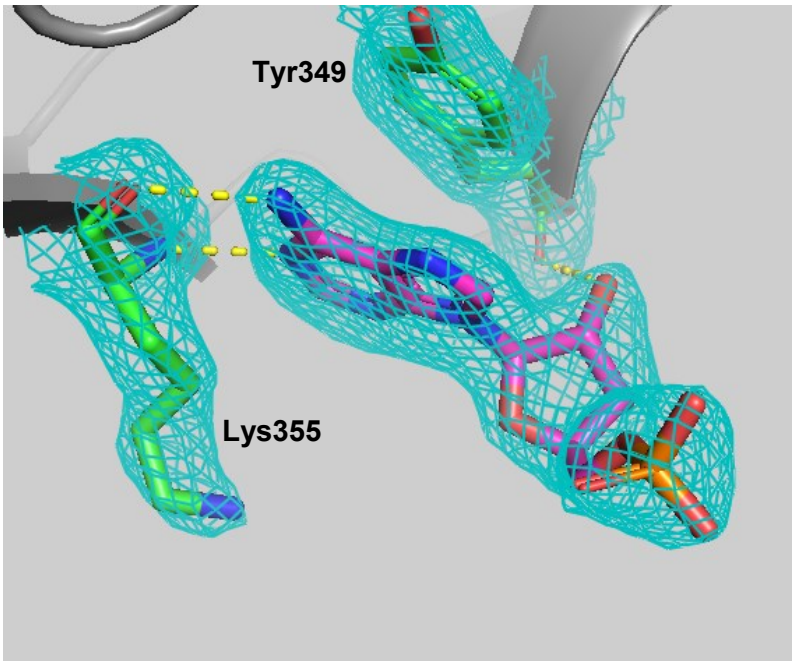

Figure 46: Free AMP in chain B.
In chain B, the 'free' AMP ligand (magenta) is seen to form two hydrogen bonds with Lys355 and once with Tyr349 and engage in pi-stacking with the latter.

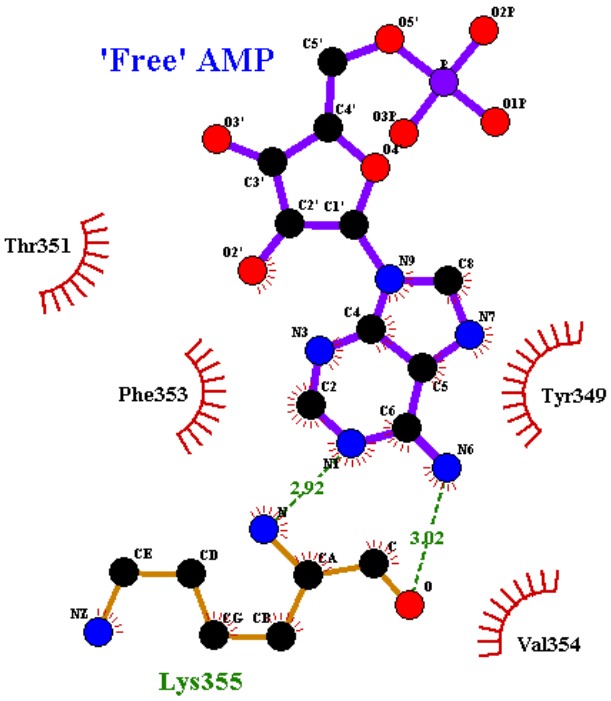

Figure 47: LAM_G4, free AMP at chain A.
This is a schematic of the free AMP in LAM_G4 chain A, prepared using LigPlot. It shows the hydrogen bonds (green dotted lines) more clearly and provides information on residues/atoms involved in hydrophobic or VDW contacts (red crescents) with the ligand.

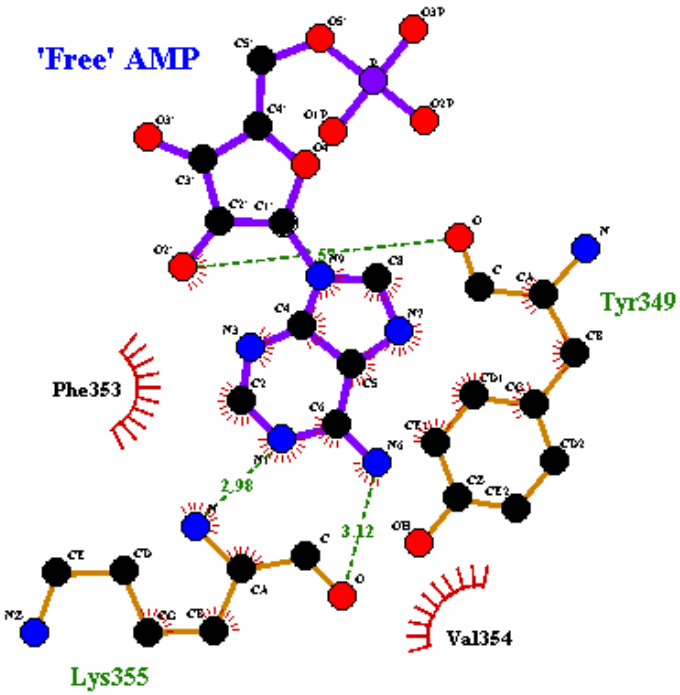

Figure 48: LAM_G4, free AMP at chain B.
This is a schematic of the free AMP in LAM_G4 chain B, prepared using LigPlot. It shows the hydrogen bonds (green dotted lines) more clearly and provides information on residues/atoms involved in hydrophobic or VDW contacts (red crescents) with the ligand.

Ultimately, after many cycles of refinement the best fit to the electron density data has been determined as AMP. First, although the density around this ligand is patchy it is connected and resembles the shape of one of the three possible ligands, and the continuity of the density makes it very unlikely to be just water in the active site. Second, despite the phosphate lacking an obvious position, it is clear that there must be at least one phosphate group present as no part of the reaction involves or produces adenosine. In lysyl-tRNA synthetase structures that contain ATP, metal ions are often found stabilising the three phosphate groups; in a structure where the intermediate is present (1E1T; *E. coli* LysRS complexed with the intermediate) they are found near the two hydrolysed phosphate groups. In the active sites of this LAM_G4 structure there are no areas of density that could represent the magnesium ions present in the crystal solution, so it is less likely to be ATP in the site.

It is possible that as more 'free AMP' is produced by the dissociation and subsequent breakdown of the lysyl-AMP intermediate AMP begins to compete with ATP for the active site. The ratio of AMP:ATP will rise as enzyme activity continues and so perhaps AMP is outcompeting ATP for the active site. To give an indication as to how tightly these free AMP molecules are bound, the program AreaIMol was used to calculate the solvent accessible surface area (SASA) buried in binding of these molecules compared to the SASA buried in binding of the AMP in the catalytic site. This is summarised in Table 17, which shows that there is almost double the surface area buried in substrate AMP binding when compared to free AMP binding. As expected due to its location on the external surface of the enzyme and as indicated by the LigPlot images, this suggests that the free AMP is bound far less tightly than the substrate AMP.

Molecule	SASA difference (Å^2)
AMP substrate in Chain A	248.1
Free AMP at Chain A	138.5
AMP substrate in Chain B	241.9
Free AMP at Chain B	133.7

Table 17: SASA differences in LAM_G4
This was calculated by running AreaIMol on LAM_G4 to calculate solvent accessible surface area, then comparing the value with those given in AreaIMol runs where each of the above molecules had been deleted.

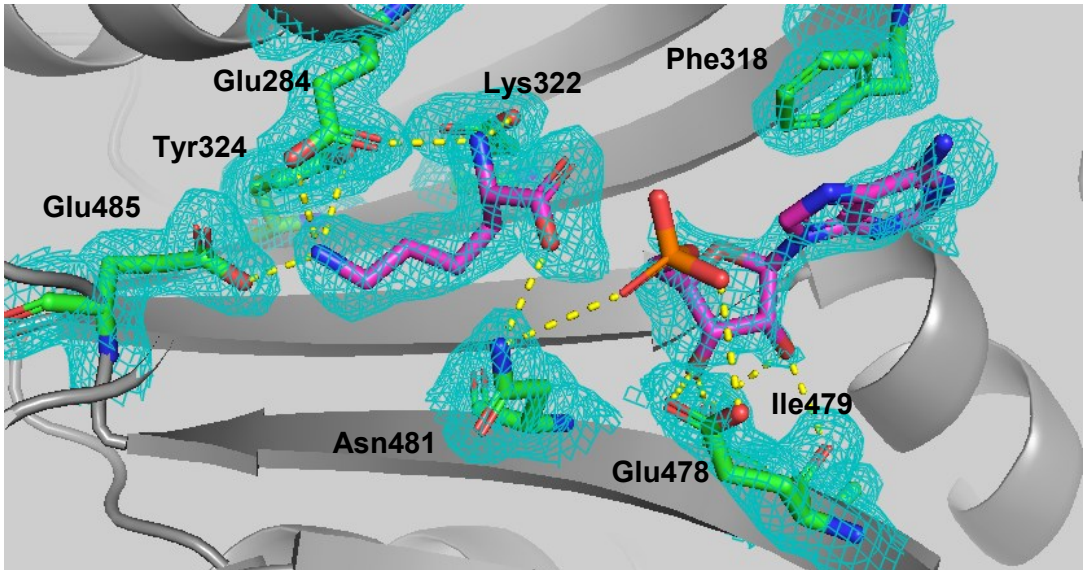

Figure 49: Active site of chain A in LAM_G4.
This is the complete active site of chain A, including both ligands – lysine and AMP. Residues that hydrogen bond with either ligand are shown with carbons coloured green, whereas the ligands are each shown with carbons coloured magenta.

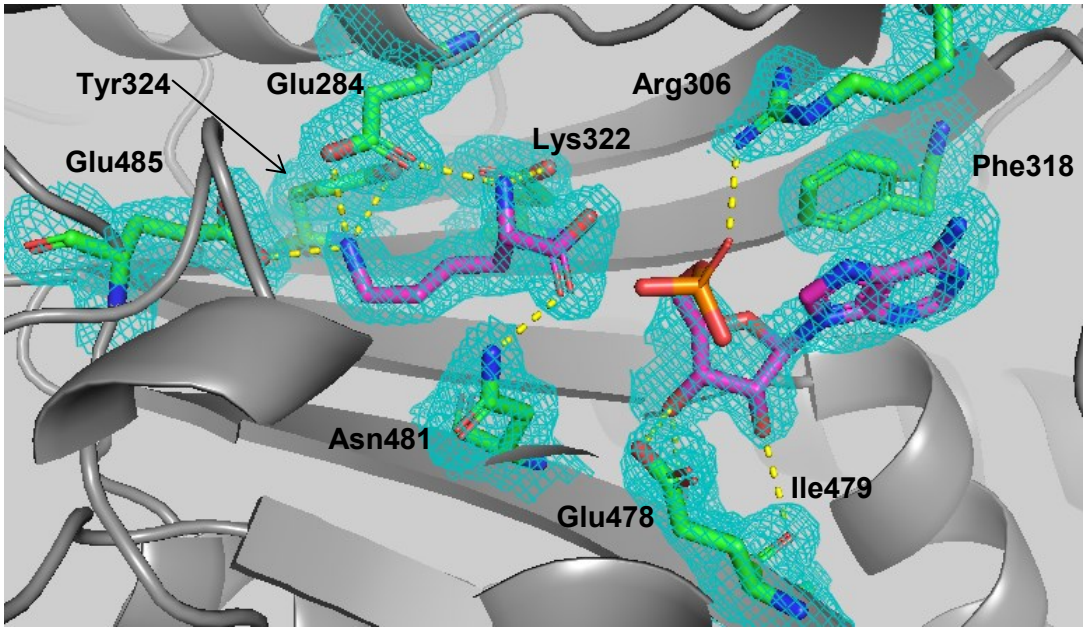

Figure 50: Active site of chain B in LAM_G4.
This is the complete active site of chain B, including both ligands, lysine and AMP. Residues that hydrogen bond with either ligand are shown with carbons coloured green, whereas the ligands are each shown with carbons coloured magenta.

The active site of chain A is seen in Figure 49, and that of chain B in Figure 50. In chain A, Glu478 and Ile479 both form hydrogen bonds with both the 2'- and 3'-hydroxyls of the AMP molecule. In chain B, only Glu478 hydrogen bonds with both and Ile479 does so with just the 3'-hydroxyl. Bearing in mind that the position of the phosphate is the most contentious part of this model, in chain A Asn481 forms a hydrogen bond with the phosphate group whereas in chain B it is Arg306 that does. That said, the phosphate group has been positioned as in other published structures with the nucleoside bent around into an approximate U-shape such that it is in an ideal position to attack the amino acid carboxylate. The adenine nucleobase pi-stacking with Phe318 seen in both chains is likely to be significant in correctly orientating the ligand.

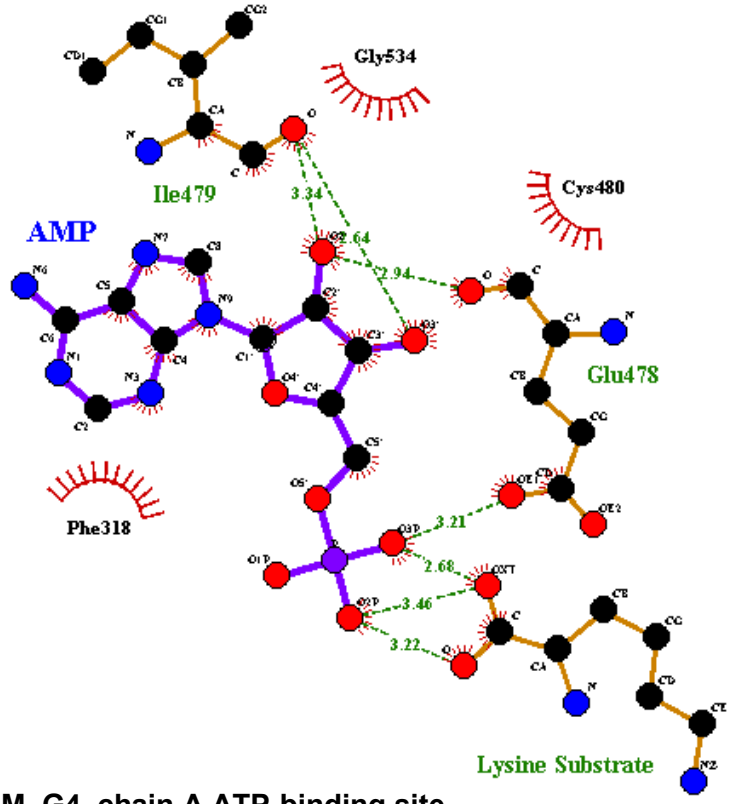

Figure 51: LAM_G4, chain A ATP-binding site.
This is a schematic of the ATP-binding site of LAM_G4 chain A, prepared using LigPlot. It shows the hydrogen bonds (green dotted lines) more clearly and provides information on residues/atoms involved in hydrophobic or VDW contacts (red crescents) with the ligand.

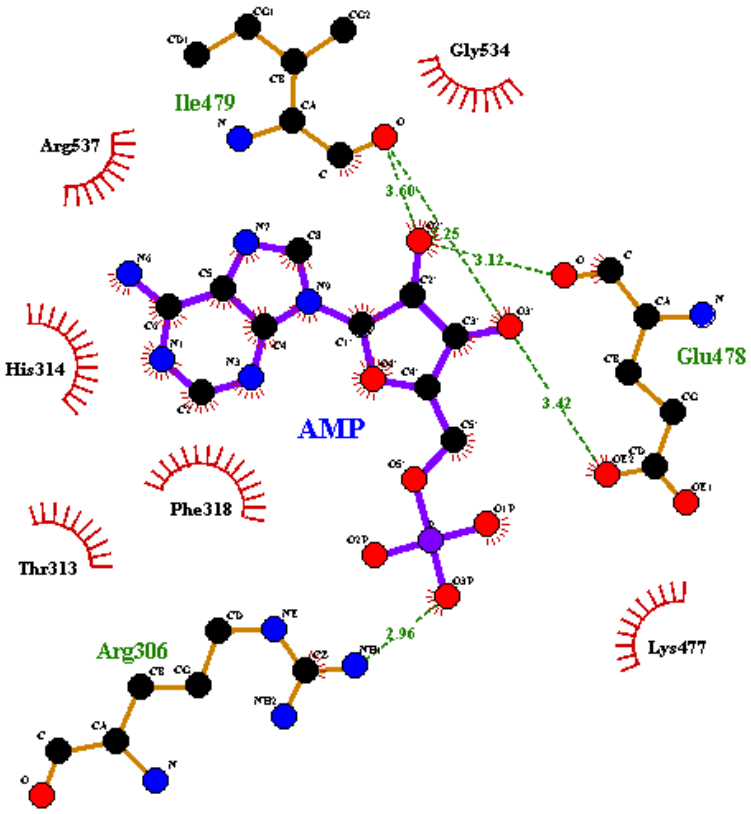

Figure 52: LAM_G4, chain B ATP-binding site.
This is a schematic of the ATP-binding site of LAM_G4 chain B, prepared using LigPlot. It shows the hydrogen bonds (green dotted lines) more clearly and provides information on residues/atoms involved in hydrophobic or VDW contacts (red crescents) with the ligand.

Figures 51 and 52 show that a number of residues are involved in hydrophobic or VDW contacts with the bound AMP ligands in each chain. These are summarised in the table below (Table 18) alongside those that contact the lysine substrate as described in figures 47 and 48 earlier.

Site	Bound ligand	Hydrophobic or VDW contacts
Chain A lysine-binding	Lysine	Ala282, Glu284, Glu322, Tyr324, Asn481, Tyr483, Glu485, Gly530, Trp531, Gly532
Chain A ATP-binding	AMP	Phe318, Glu478, Ile479, Cys480, Gly534
Chain B lysine-binding	Lysine	Ala282, Glu284, Glu322, Tyr324, Asn481, Tyr483, Glu485, Gly530, Trp531, Gly532
Chain B ATP-binding	AMP	Arg306, Thr313, His314, Phe318, Lys477, Glu478, Ile479, Gly534, Arg537

Table 18: Summary of hydrophobic or VDW contacts in LAM_G4 catalytic sites.
This table lists the residues that are involved in hydrophobic or VDW contacts with the bound substrates in each catalytic site as reported by LigPlot.

Comparison with 3BJU

In 3BJU, the structure of a tetrameric human lysyl-tRNA synthetase with ATP and lysine bound to each chain, the ribose forms hydrogen bonds with an adjacent Glu / Ile pair as well (positions 494 and 495, respectively, in that structure) but the adenine ring also forms two hydrogen bonds with an asparagine residue (position 332 in 3BJU) of which the equivalent (position 315 in LAM_G4) is just slightly too far away in the LAM_G4 model. It is worth noting that in 3BJU the four different active sites do not have exactly the same non-covalent interactions with their ligands, indicating that there is some flexibility in the binding interaction. Figure 53 shows the chain A catalytic site of LAM_G4 again but this time also highlighting the residues that are involved in hydrogen bonding in 3BJU but not in LAM_G4.

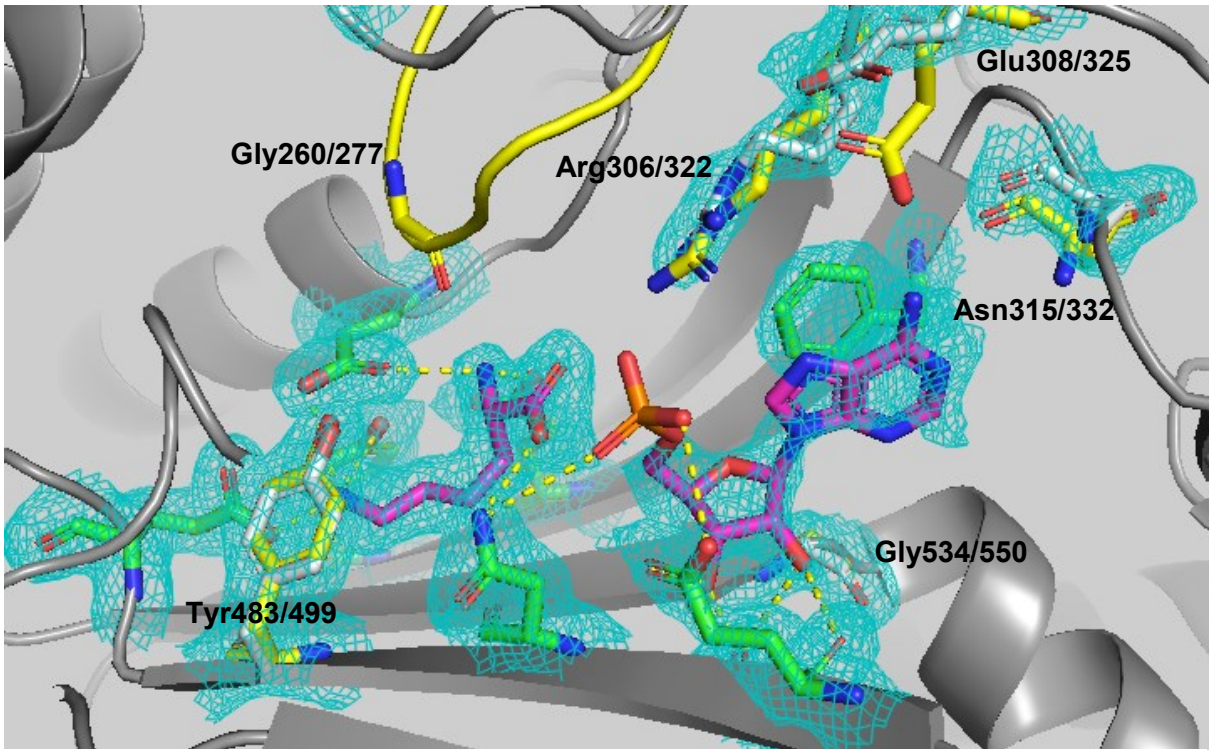

Figure 53: Active site comparison of LAM_G4 chain A and 3BJU chain A.
This image shows the chain A active site of LAM_G4. Residues in LAM_G4 that hydrogen bond (yellow dashed line) with the ligands are shown in green. Also highlighted are those that hydrogen bond with the ligands in 3BJU (yellow) but not in LAM_G4 (white). The density map belongs to LAM_G4, to show that there is little ambiguity in the positions of these residues. Residue numbers in labels are set out as "<Residue>LAM_G4/3BJU".

In Figure 53, it can be seen that most residues that hydrogen bond to the ligands in LAM_G4 also do so in 3BJU – but there are some differences. Firstly, there are residues that are just slightly too far away – Gly534 and Tyr483, both only 3.6 Å away – that could play a role in ligand binding but that do not in this model. The more significant differences are the comparatively larger distances between the phosphate group of AMP and Arg306 (4.2 Å) and Asn315 (4.8 Å). A limitation is that the positioning of the phosphate is potentially inaccurate given the lack of clear density, so the former of those differences may not be significant.

One major difference is that, as in LAM_H2, the Glu308 side chain is twisted away from the active site and is therefore unlikely to be involved in catalytic activity. This represents a major deviation between the mammalian and *T. congolense* enzymes.

The second major deviation is, again, the position of Loop-Alpha. Loop-Alpha is in conformation 1 in 3BJU, with Gly277 hydrogen bonding with the lysine, yet in LAM_G4, that loop is in conformation 2 and thus the corresponding Gly260 is much further away. This is illustrated in Figure 54.

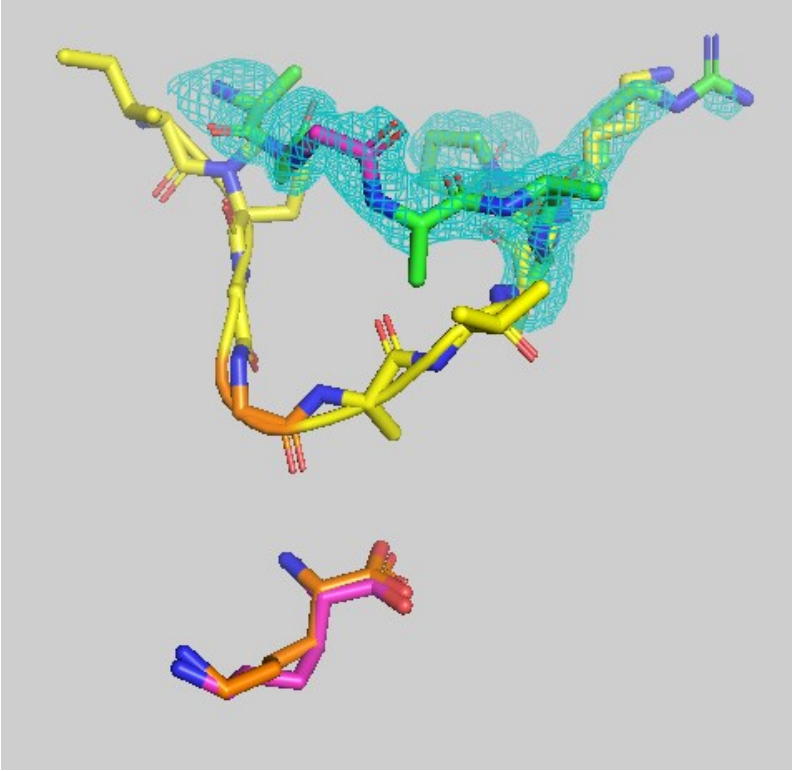

Figure 54: Loop-Alpha comparison of LAM_G4 chain A and 3BJU chain A.
This image shows the different conformation of Loop-Alpha in LAM_G4 compared to 3BJU. 3BJU residues are shown with yellow carbons, with Gly277 and the lysine ligand in orange, whereas LAM_G4 residues have green carbons with Gly260 and the lysine ligand in magenta.

LCM_F3

Unlike the two previous structures that were produced from ATP-containing crystallisation experiments, LCM_F3 was produced using cordycepin, a nucleoside analogue, instead of ATP. The LCM_F3 crystal structure contains homodimeric LysRS1 with both active sites occupied by the potential inhibitor cordycepin. Figure 55 shows a cartoon rendering of the entire structure, featuring the same overall fold that was previously described, coloured by chain as before.

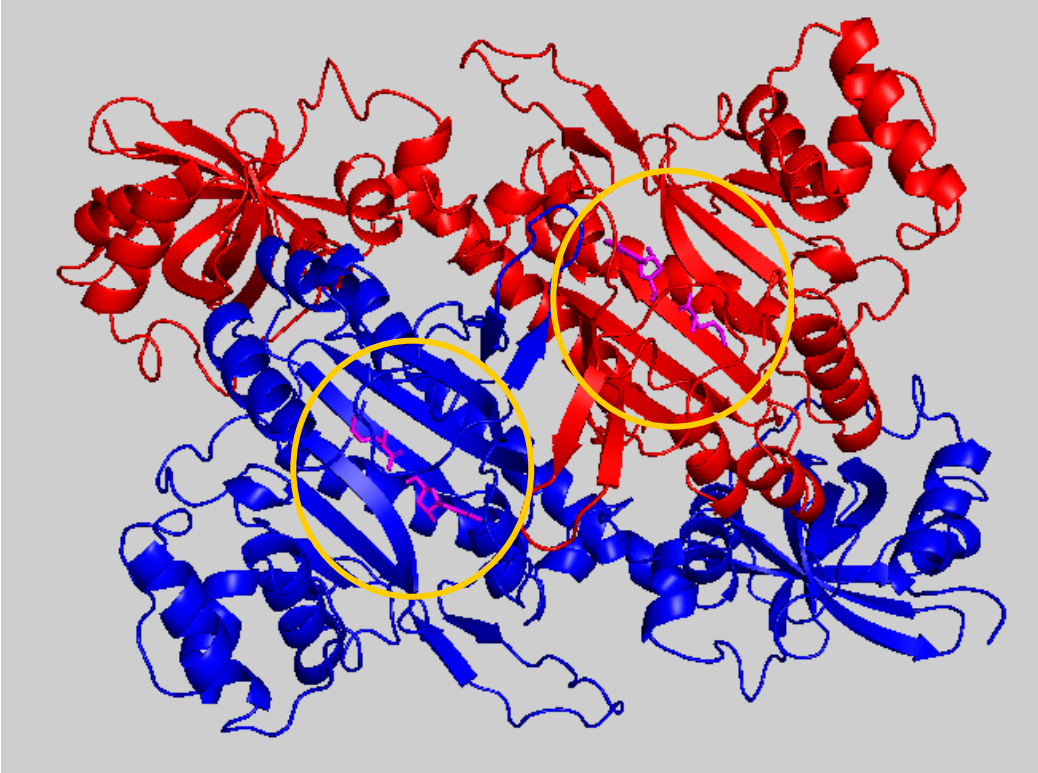

Figure 55: An overview of LCM_F3.
This image shows an overview of the LCM_F3 structure. Chain A is in red and chain B is in blue, whereas the ligands are shown in magenta. The active sites are highlighted with a yellow ring.

In order to highlight any major structural differences between the monomers, Figure 56 shows the C-alpha traces of both monomers overlaid and coloured by RMSD.

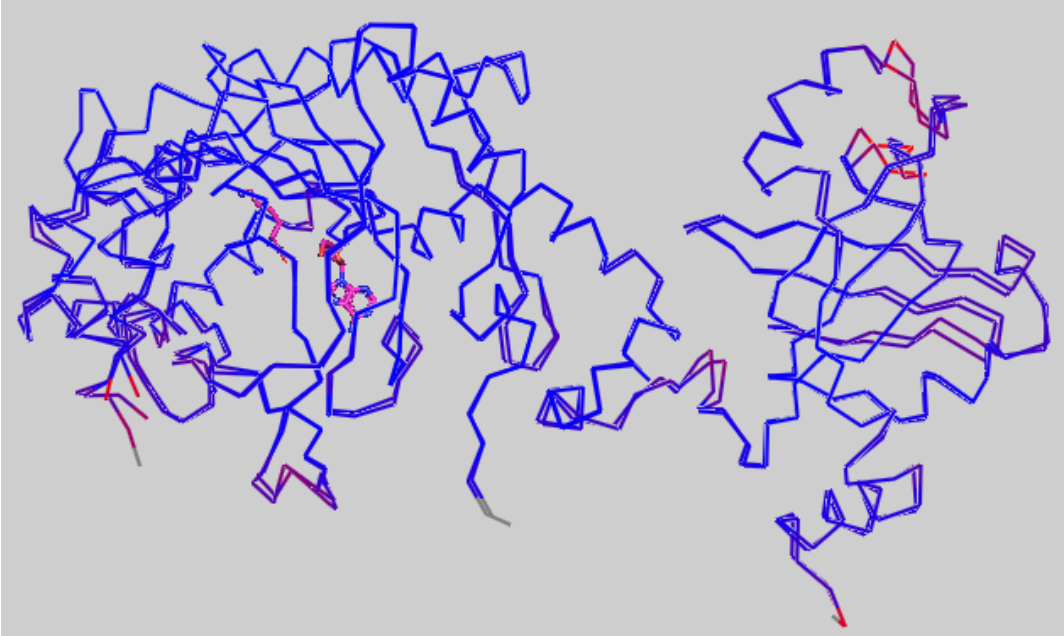

Figure 56: C-alpha trace overlay coloured by RMSD (LCM_F3_A v LCM_F3_B). This figure shows the C-alpha traces of chain A and chain B of LCM_F3. The overall RMSD for this alignment is 0.34. A PyMOL script was used to align the two and then colour them by local RMSD. A rainbow spectrum of colours indicates relative RMSD, starting at a dark blue for minimum RMSD ranging down to red for maximum RMSD. Grey indicates areas that were excluded from the alignment. Cordycepin is visible to show the location of the catalytic site.

LCM_F3 shows that this analogue is present in the active site with excellent representation in the electron density and therefore is likely to act, at least to some degree, as a competitive inhibitor to *T. congolense* lysyl-tRNA synthetase I.

In chain A, Figure 57 shows Tyr324, Glu485, Tyr483 and Glu284 hydrogen bonding with lysine at the epsilon-amino group. Glu284 additionally forms a hydrogen bond with the alpha-amino group along with Glu322, whereas Asn481 and Arg306 both do so with the carboxyl group of the amino acid.

Cordycepin can be seen to hydrogen bond with Arg537, Glu478 and Ile479 via the only hydroxyl group of its ribose moiety, and with Arg306 via a nitrogen group within the purine ring section.

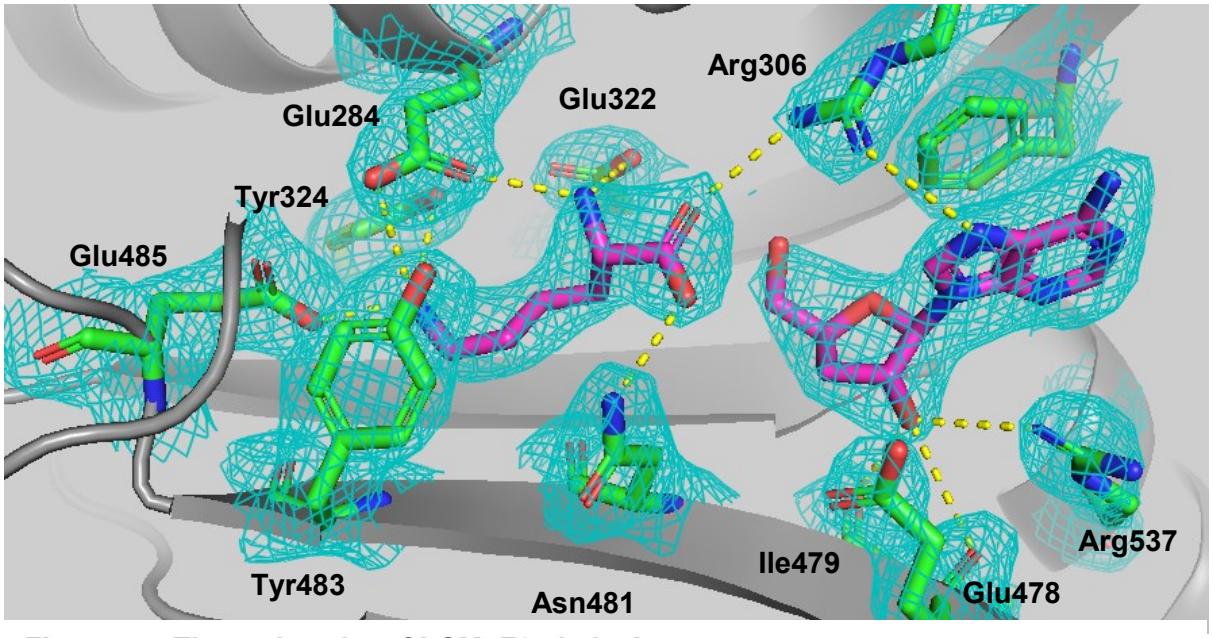

Figure 57: The active site of LCM_F3 chain A.
This image shows the ligands lysine and cordycepin, both coloured with magenta carbons, in the active site of chain A. Catalytic residues are coloured with green carbons.

Figures 58 and 59 show this same area visualised in Ligplot so that the bonds can be more clearly seen, as well as to give an insight into which residues engage in hydrophobic or VDW contacts with the bound ligands.

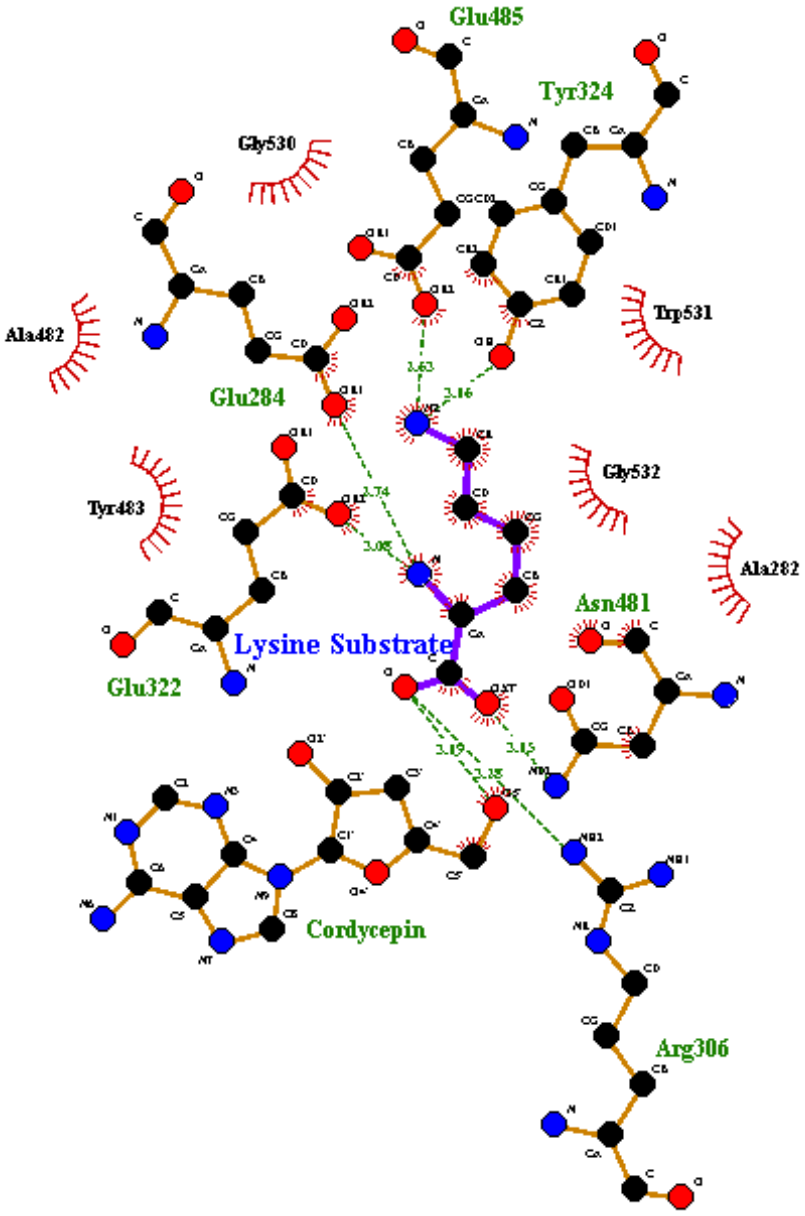

Figure 58: LCM_F3 chain A lysine-binding site in LigPlot.
This is a schematic of the chain A lysine-binding site of LCM_F3, prepared using LigPlot. It shows the hydrogen bonds (green dotted lines) more clearly and provides information on residues involved in hydrophobic or VDW contacts (red crescents) with the bound ligand.

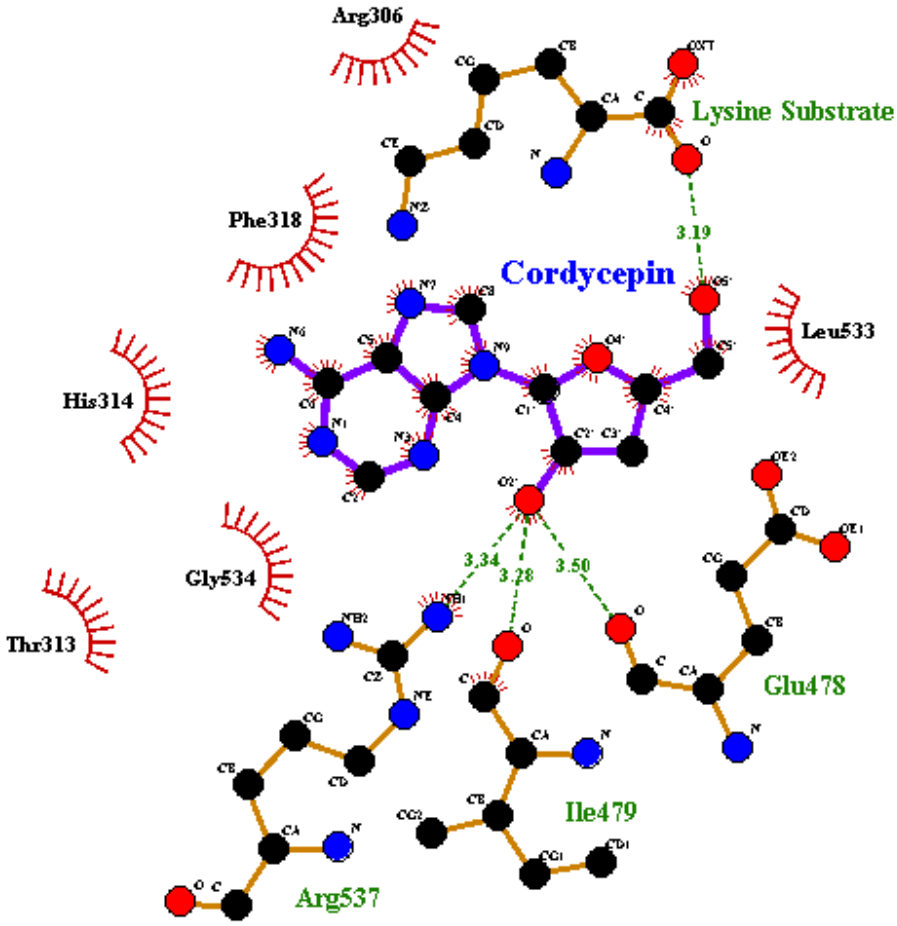

Figure 59: LCM_F3 chain A ATP-binding site in LigPlot.
This is a schematic of the chain A ATP-binding site of LCM_F3, prepared using LigPlot. It shows the hydrogen bonds (green dotted lines) more clearly and provides information on residues involved in hydrophobic or VDW contacts (red crescents) with the bound cordycepin ligand.

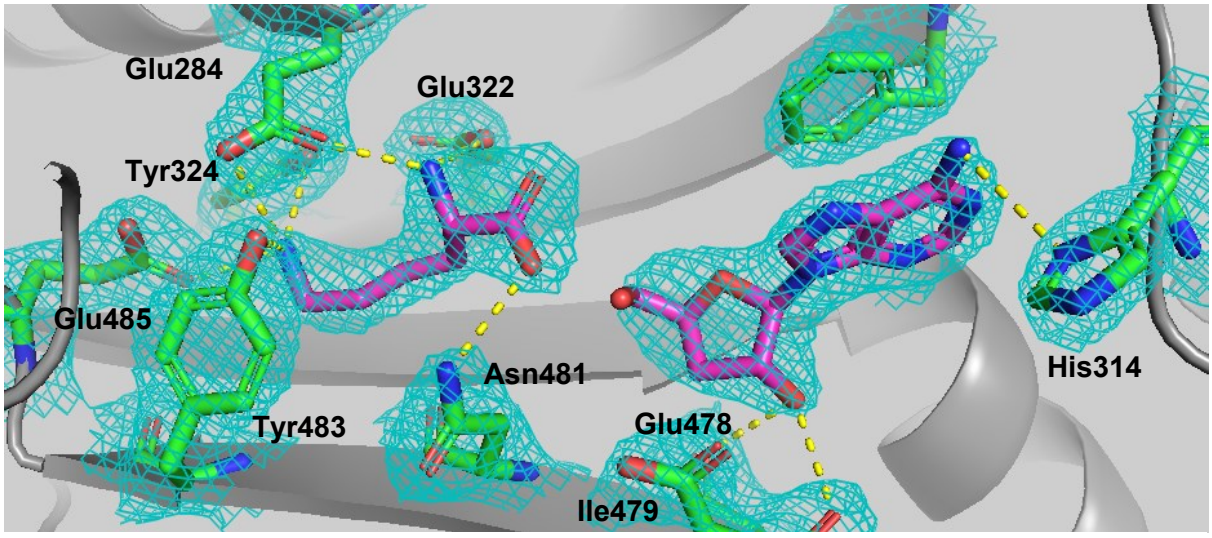

Figure 60: The active site of LCM_F3 chain B.
This image shows the ligands lysine and cordycepin, both coloured with magenta carbons, in the active site of chain B. Catalytic residues are coloured with green carbons.

Chain B has the lysine ligand forming hydrogen bonds with Tyr324, Glu485, Tyr483 and Glu284 again at the epsilon-amino group, with Glu284 and Glu322 hydrogen bonding to the alpha amino group. Asn481 is the only residue to hydrogen bond with the carboxyl group in this second chain. These are seen in figures 60-62.

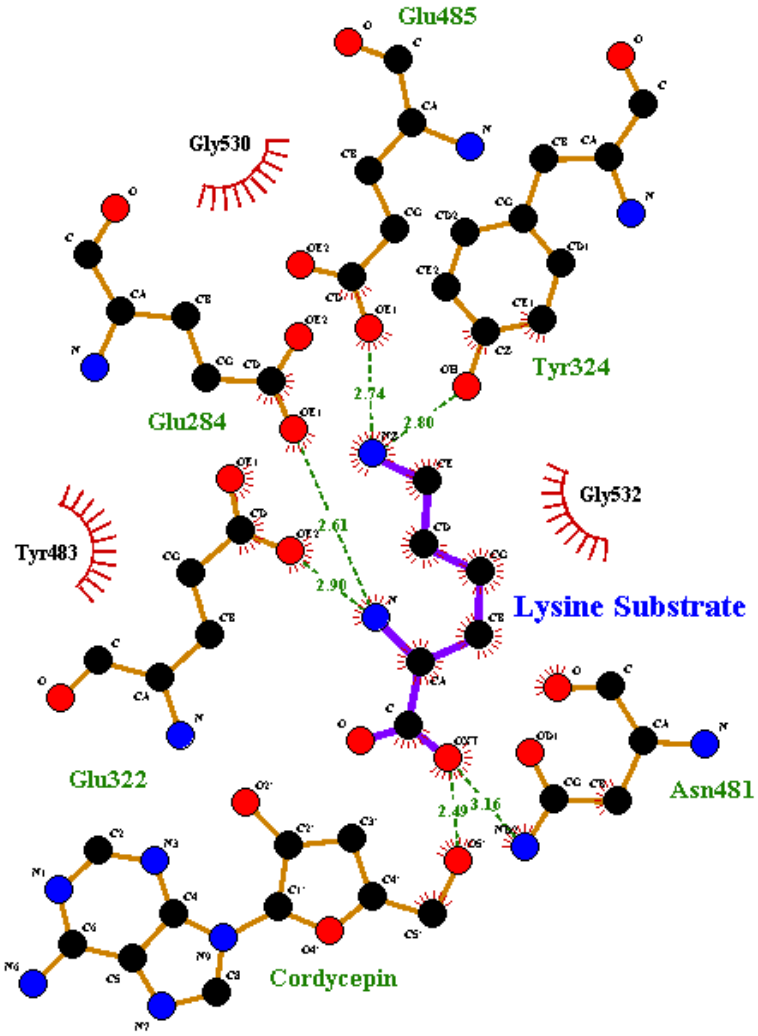

Figure 61: LCM_F3 chain B lysine-binding site in LigPlot.
This is a schematic of the chain B lysine-binding site of LCM_F3, prepared using LigPlot. It shows the hydrogen bonds (green dotted lines) more clearly and provides information on residues involved in hydrophobic or VDW contacts (red crescents) with the bound ligand.

Cordycepin forms hydrogen bonds with Ile479 and Glu478 via its lone ribose hydroxyl group and, unusually, His314 via the purine amino group. This hydrogen bond with His314 is not assigned by LigPlot. It is thought that, in principle, a (weak) hydrogen bond is possible due to the capability of aromatic rings to act as hydrogen bond acceptors. The nature of this bond remains unclear given that the crystal pH (6.5) is close to histidine pI (7.59), which will affect its strength. As this bond is not present in the chain A active site (due to His314 orientation), it is reasonable to assume that it is not required for cordycepin binding.

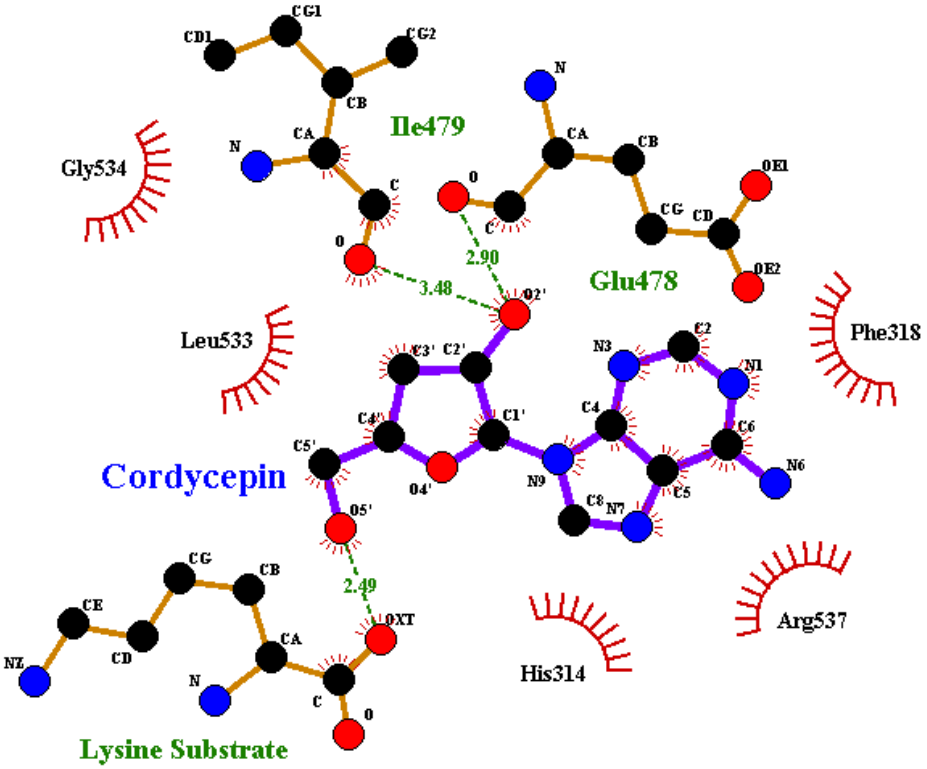

Figure 62: LCM_F3 chain B ATP-binding site in Ligplot.
This is a schematic of the chain B ATP-binding site of LCM_F3, prepared using Ligplot. It shows the hydrogen bonds (green dotted lines) more clearly and provides information on residues involved in hydrophobic or VDW contacts (red crescents) with the bound cordycepin ligand.

As expected, cordycepin binds in a similar fashion to the ATP that it is mimicking though with fewer non-covalent bonds. There are not as many hydrogen bonds, particularly because the cordycepin is lacking the additional hydroxyl group and the phosphates that should both form additional non-covalent bonds. In addition to the hydrogen bonds already discussed, there are a number of residues involved in hydrophobic or VDW contacts with the bound ligands. These are summarised in Table 19 below.

Site	Bound ligand	Hydrophobic or VDW contacts
Chain A lysine-binding	Lysine	Ala282, Glu284, Glu322, Tyr324, Asn481, Ala482, Tyr483, Glu485, Gly530, Trp531, Gly532
Chain A ATP-binding	Cordycepin	Arg306, Thr313, His314, Phe318, Ile479, Leu533, Gly534, Arg537
Chain B lysine-binding	Lysine	Glu284, Glu322, Tyr324, Asn481, Tyr483, Glu485, Gly530, Gly532
Chain B ATP-binding	Cordycepin	His314, Phe318, Glu478, Ile479, Leu533, Gly534, Arg537

Table 19: Summary of hydrophobic or VDW contacts in LCM_F3 catalytic sites. This table lists the residues that are involved in hydrophobic or VDW contacts with the bound substrates in each catalytic site as reported by Ligplot.

Interestingly, there are no 'free cordycepin' molecules bound to either chain as there were in the other two structures at Tyr349. This may be because cordycepin does not have the 3'-hydroxyl group that forms the hydrogen bond with Tyr349, despite containing the ring that would enable pi-stacking and, potentially, the hydrogen bond with Lys355.

Comparison with 4YCU

4YCU is a structure that shows cladosporin, another nucleoside analogue, in complex with the human LysRS. Although cladosporin is not as similar to cordycepin as cordycepin is to ATP, sharing only the ring structure and general dimensions that allow them to mimic ATP, they are being used for the same purpose: as competitive inhibitors that occupy the ATP-binding site to reduce aminoacyl-tRNA synthetase activity. Cladosporin has already been shown to target the LysRS of *Plasmodium falciparum* (Hoepfer et al., 2012), and so it may be useful to examine how cladosporin binds to the mammalian LysRS compared to how cordycepin binds to *T. con* LysRS1. The structure of cladosporin is shown in Figure 63.

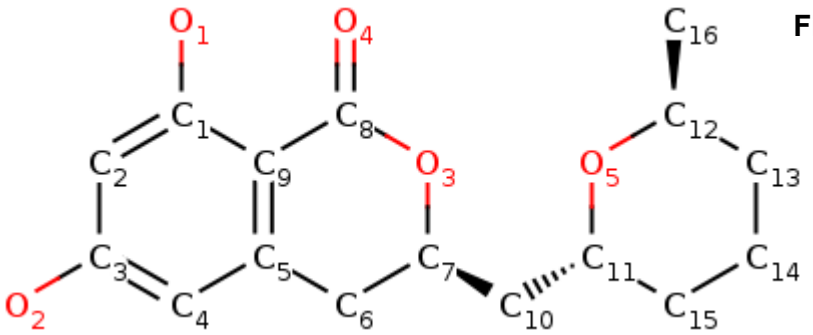

Figure 63: Cladosporin structure.

In Figure 64, this comparison is visualised in PyMOL. It can be seen that cladosporin forms only three hydrogen bonds with the mammalian LysRS and that two of these involve the same residue Asn332. The equivalent residue Asn315 in LCM_F3 is slightly too far away from the cordycepin to

form any similar hydrogen bonds. The other hydrogen bond between cladosporin and the enzyme in 4YCU is Glu325 where the equivalent in LCM_F3, Glu308, is one of the consistent differences observed in this study between mammalian and *T.congolense* LysRSes. As in all the others, Glu308 is twisted upward and so cannot bind to the ligand. Cladosporin is actually positioned differently within the 4YCU binding pocket compared to cordycepin in LCM_F3, which more closely imitates the natural ligands in terms of positioning, and this leads to substantial differences in hydrogen bonding between the two analogues. It indicates that, as expected, the most significant feature to retain when thinking of a molecule to bind to this site is the double-ring structure that engages in pi-stacking interactions with Phe318.

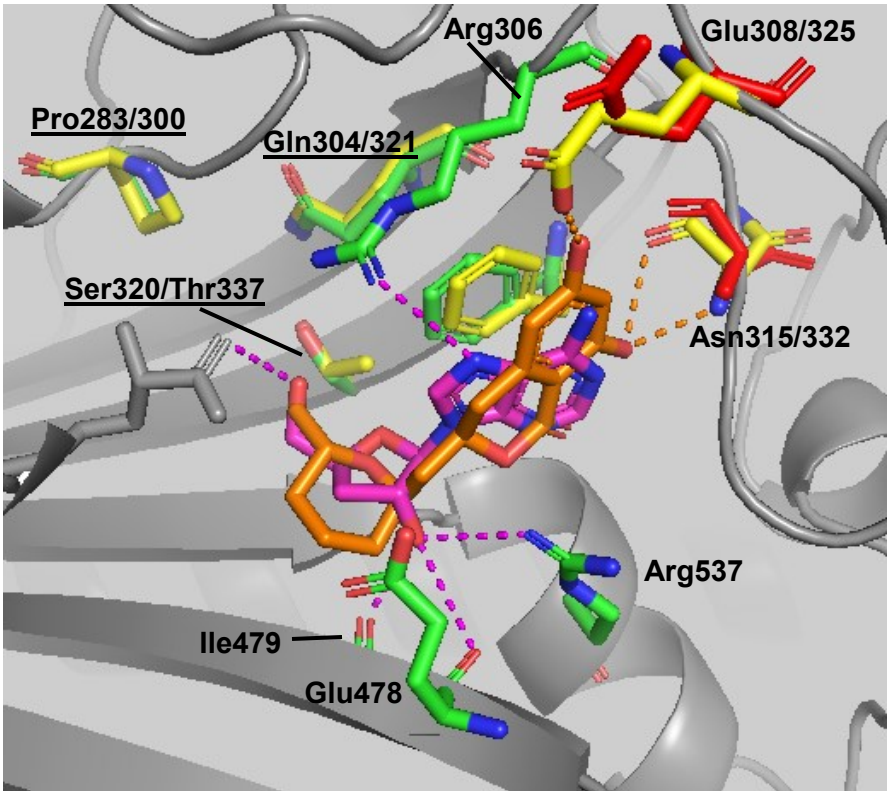

Figure 64: The active site of LCM_F3 showing residues of interest from 4YCU overlaid.
LCM_F3 residues are coloured with green carbons and the cordycepin ligand is coloured with magenta carbons, whereas 4YCU residues are coloured with yellow carbons and cladosporin is coloured with orange carbons. Red residues are those that hydrogen bond with cladosporin in 4YCU but not with cordycepin in LCM_F3. Residue labels are set out in a [Residue]LCM_F3/4YCU format and underlined residues are those identified as having effect on cladosporin specificity.

The underlined labels in Figure 64 indicate three residue positions that have been identified previously as determining specificity for cladosporin; in the mammalian LysRS, these are positions 300, 321 and 337 (283, 304 and 320 in *T. congolense* LysRS1). The study from which 4YCU originates (Fang et al., 2015) explains that in other aaRSes examined, the equivalent residues either have larger or shorter side chains and cannot form stable interactions with cladosporin – in

this manner, they provide specificity to cladosporin binding. These three residues are said to be highly variable within and between different aaRS families which makes it unlikely for them to be important to the aminoacylation reaction. This can be seen in the figure above in that they are too far from cladosporin in 4YCU – and from cordycepin in LCM_F3 – to be directly involved in ligand binding interactions. This is important because it is a clear indication that residues not involved with the actual reaction can determine how specifically a proposed (or designed) inhibitor can bind to this pocket and affect this system. LigPlot was used to include hydrophobic or VDW contacts in the comparison, the results of which are seen in Figure 65 and Table 20.

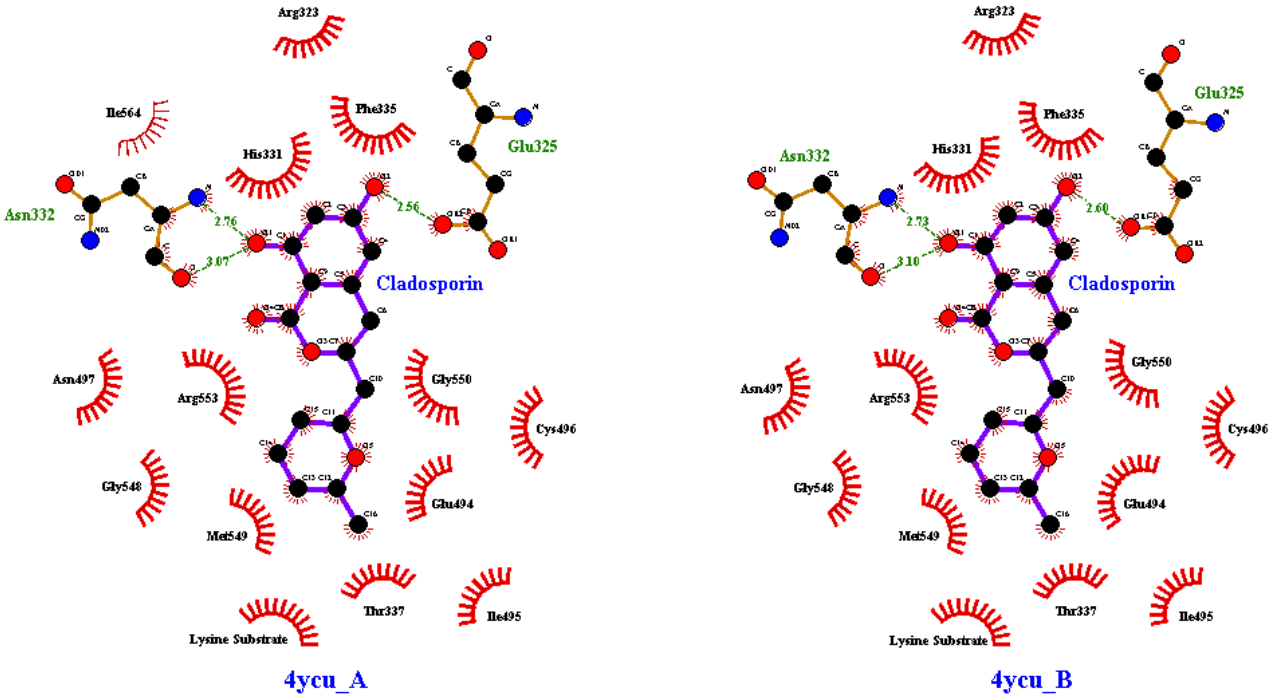

Figure 65: ATP-binding sites of 4YCU chain A (left) and chain B (right).
This is a schematic of both ATP-binding sites in 4YCU, prepared using Ligplot. It shows the hydrogen bonds (green dotted lines) more clearly and provides information on residues involved in hydrophobic or VDW contacts (red crescents) with the bound cladosporin ligands.

Structure	Chain	Hydrophobic or VDW contacts
LCM_F3	A	Arg306, Thr313, His314, Phe318, Ile479, Leu533, Gly534, Arg537
LCM_F3	B	His314, Phe318, Glu478, Ile479, Leu533, Gly534, Arg537
4YCU	A	**Arg323(306),** Glu325(308), **His331(314),** Asn332(315), **Phe335(318),** Thr337(Ser320), **Glu494(478), Ile495(479),** Cys496(480), Asn497(481), Gly548(532), **Met549(Leu533), Gly550(534), Arg553(537),** Ile564(548),
4YCU	B	**Arg323(306),** Glu325(308), **His331(314),** Asn332(315), **Phe335(318),** Thr337(Ser320), **Glu494(478), Ile495(479),** Cys496(480), Asn497(481), Gly548(532), **Met549(Leu533), Gly550(534), Arg553(537),**

Table 20: Summary of hydrophobic or VDW contacts in LCM_F3 and 4YCU ATP-binding sites.
This table lists the residues that are involved in hydrophobic or VDW contacts with the ligands bound at the ATP-binding sites in both LCM_F3 and 4YCU. In the 4YCU rows, the equivalent LCM_F3 residues are shown in brackets. The bolded residues are the 4YCU residues involved whose equivalent LCM_F3 residues are also involved.

Almost double the number of residues in 4YCU form hydrophobic or VDW contacts with cladosporin than residues in LCM_F3 do with cordycepin, but this is unsurprising given the relative sizes of the two ligands. Cladosporin is much larger due to containing three larger rings and therefore will be within range of more residues. Table 21 details the results of SASA calculations performed on these two structures in order to give some indication of how contacts might influence how tightly each ligand is bound in its respective structure.

Molecule	SASA difference (Å^2)
4YCU Chain A Cladosporin	223.0
4YCU Chain B Cladosporin	217.4
LCM_F3 Chain A Cordycepin	219.0
LCM_F3 Chain B Cordycepin	218.9

Table 21: SASA differences in 4YCU and LCM_F3 with regards to ATP-binding site.
These were calculated by running AreaIMol on each structure to calculate solvent accessible surface area, then comparing the value with those given in AreaIMol runs where each of the above ligands had been deleted.

These results indicate that both ligands form similar total contact areas with their respective structures, which may suggest a similar strength of association. This is despite cladosporin forming more contacts with surrounding residues. It is likely that these additional contacts are simply a

result of its larger size, therefore requiring more residues involved to achieve the same level of overall contact as the smaller cordycepin.

Molecule	SASA difference (Å²)
3BJU Chain A ATP	259.8
3BJU Chain B ATP	261.1
LAM_G4 Chain A AMP substrate	248.1
LAM_G4 Chain B AMP substrate	241.9
LCM_F3 Chain A Cordycepin	219.0
LCM_F3 Chain B Cordycepin	218.9

Table 22: SASA differences in the ATP-binding sites of 3BJU, LAM_G4 and LCM_F3.
These were calculated by running AreaIMol on each structure to calculate solvent accessible surface area, then comparing the value with those given in AreaIMol runs where each of the above molecules had been deleted.

Table 22 shows the SASA differences in the ATP-binding sites of chain A and chain B in structures 3BJU, LAM_G4 and LCM_F3. Although roughly the same solvent-accessible surface area is buried in cordycepin binding as in cladosporin binding, there is notably more buried during AMP binding – and more again during the binding of ATP to mammalian LysRS in 3BJU. This is expected given that ATP is the 'true' binding ligand. AMP is not part of the lysyl-AMP intermediate has no role to play in the reaction other than being released.

The next figures compare the chain of LAM_H2 containing the lysyl-AMP intermediate with 3BJU and an LCM_F3 chain.

Comparisons with LAM_H2 and 3BJU

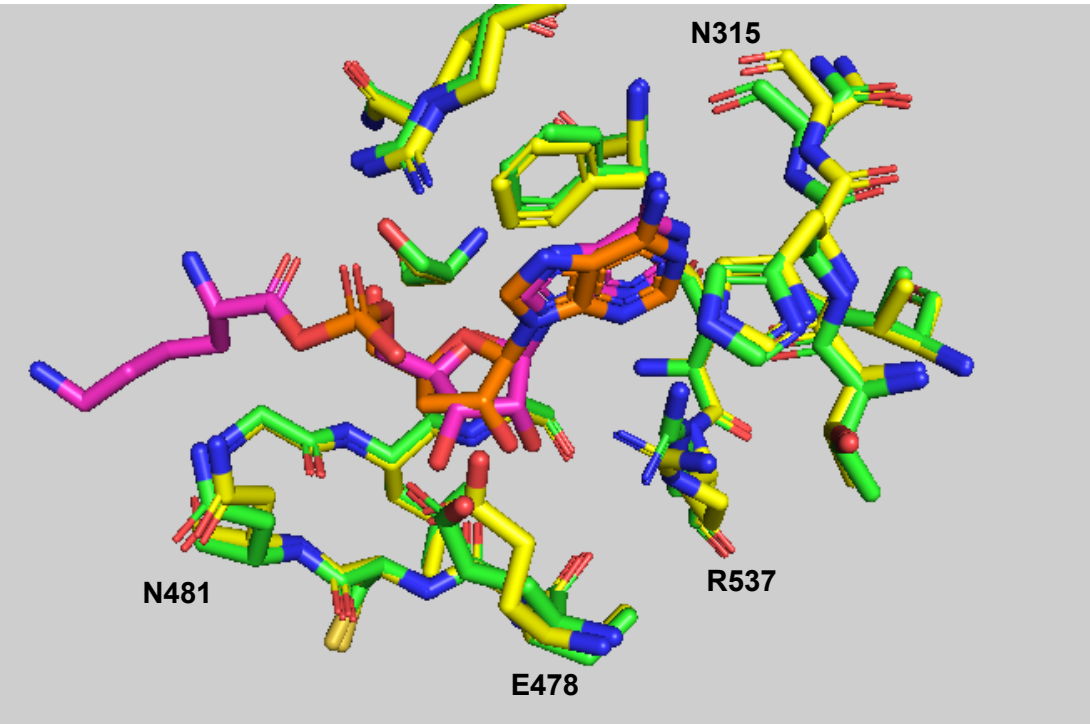

Figure 66: A comparison between the binding pockets of LCM_F3 and LAM_H2.
LCM_F3 is shown with green carbons and a magenta ligand, whereas LAM_H2 is shown with yellow carbons and an orange ligand. The residues were identified through the use of the 'surface map' feature in PyMOL – all residues forming part of the surface exposed to the ligand were selected. Labelled residues are those with different positions.

Figure 66 is a visual comparison between the binding pockets of LCM_F3 and LAM_H2, showing the differences in the pockets – Arg537, Glu478, Asn481 and Asn315. Glu478 and Asn315 have been previously mentioned as catalytic residues in some structures, yet Asn481 and Arg537 have not been seen to form any direct contacts with substrates and therefore are unlikely to have direct influence on the aminoacylation reaction. These four residues show how the actual shape of the binding pocket itself changes when cordycepin binds and so could be worth comparing with bovine LysRS; if they are different, they could provide a way for cordycepin (or a similar, hypothetical designed drug) to distinguish between the parasite and host enzymes. Another comparison is made in Figure 67, but this time between 3BJU and LCM_F3, to indicate the differences between the mammalian and *T. congolense* enzyme sites. Also included in this figure is Glu308/325, as it is ligand-binding in 3BJU.

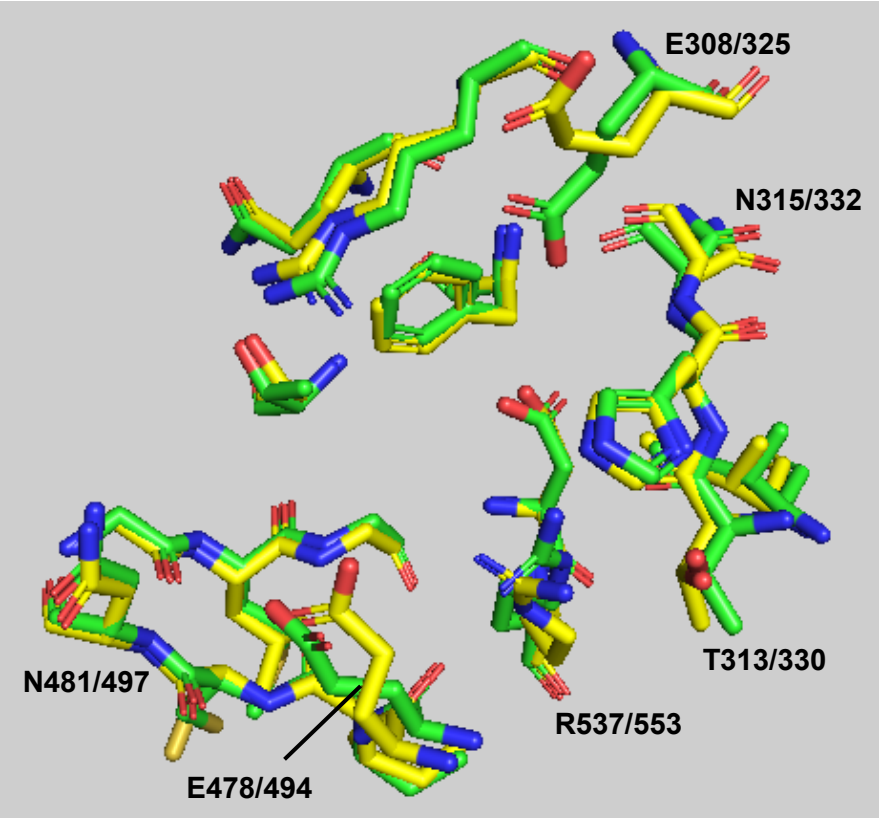

Figure 67: A comparison between the binding pockets of 3BJU and LCM_F3. 3BJU is shown with green carbons, whereas LCM_F3 is shown with yellow carbons. The residues were identified through the use of the 'surface map' feature in PyMOL – all residues forming part of the surface exposed to the ligand were selected. Labelled residues are those with different positions. The ligands have been excluded from this image so as to be able to see the residues more clearly.

In Figure 68, another such comparison can be seen but this time with the surface maps of each pocket displayed. This enhances the difference that the conformation of Glu308/325 makes and is therefore very likely to be significant when it comes to specificity of any potential drugs to the *T. congolense* LysRS1 active site. An additional group stretching upward into that space, which is blocked off by Glu325 in the mammalian enzyme structure but made accessible by the absence of Glu308 there in the *T. congolense* enzyme structure, could potentially allow a molecule to bind to the parasitic enzyme active site but be unable to fit into the site of the host enzyme.

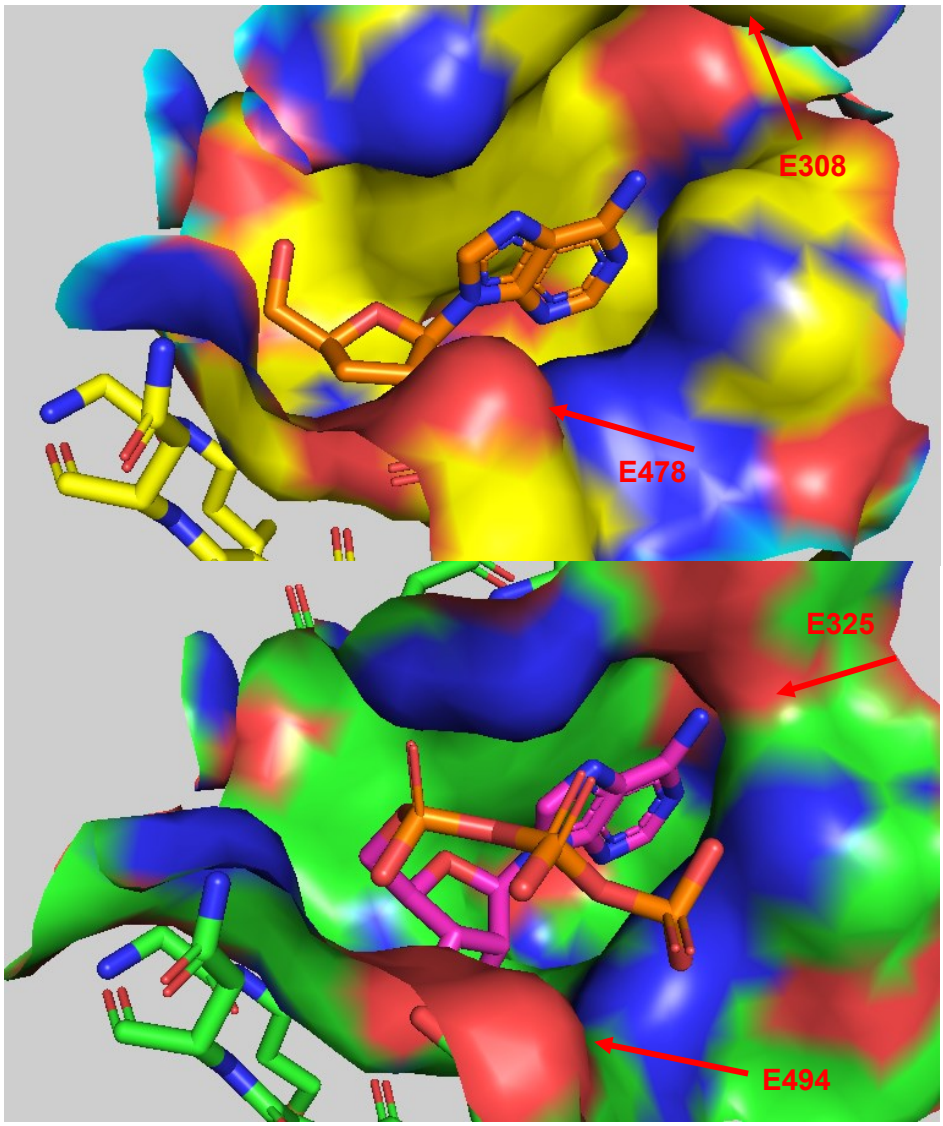

Figure 68: Differences in the binding pockets of 3BJU and LCM_F3.
LCM_F3 is shown in the top image, whereas 3BJU is shown in the bottom one. Red arrows indicate the major, most visible differences between the two.

Mentioned earlier in this study, there is an unvalidated bovine LysRS sequence (F1MMK8) available on Uniprot. Because it is unreviewed, a very similar sequence – that of human LysRS – has been mainly used for comparisons. However, it stands to reason that the active site would be the most conserved part of any enzyme, and therefore between two sequences with high similarity overall they are very likely to be near identical. A homology model for bovine LysRS was prepared using the 'SWISS-MODEL' online resource and was briefly consulted to confirm that the glutamate at position 308 in the trypanosome sequence and position 325 in the human sequence is conserved in the bovine sequence at position 351. The homology model places it in the same conformation as the mammalian, as would be expected. However, at this stage the unreviewed bovine sequence is insufficiently validated to draw any major assertions other than showing that the residue is conserved in the best currently-available sequence of bovine LysRS.

Conclusions & Future Directions

Comparing all three of the structures solved in this study, as well as making comparisons with the two provided in the preceding study, allows for a comprehensive evaluation of which residues in the active site are involved in ligand binding. The RMSD values of the alignments between each of the structures in their entirety is summarised below in Table 23.

Alignment	RMSD (Å)
Existing – LCM_F3	0.39
Existing – LAM_H2	0.42
Existing – LAM_G4	0.28
LCM_F3 – LAM_H2	0.39
LCM_F3 – LAM_G4	0.21
LAM_H2 – LAM_G4	0.34
LCM_F3 A – B	0.34
LAM_H2 A – B	0.53
LAM_G4 A – B	0.30

Table 23: Compiled alignment RMSD values.
This figure shows RMSD values of the main alignments.

These values are consistent with the comments made earlier in the text and with what is contained in each active site. For example, it makes sense that LCM_F3 (Cordycepin) and LAM_G4 (AMP) are more closely aligned than either of them with LAM_H2 (Intermediate) given the structural similarities between cordycepin and AMP. There are no major, consistent structural deviations in any of the structures solved in this report compared to the existing lysine-complexed structure, except for the loops already discussed. The largest alignment discrepancy as seen in this table is between the two chains of LAM_H2, but this is presumably just a result of this being the only structure with different active site contents in each chain (one chain containing the intermediate and the other containing lysine alone). This would be expected to have some impact on enzyme structure – such as Loop-Alpha being in a different conformation in each of the monomers in that structure.

Residues identified that form hydrogen bonds with lysine, in chains that do not contain the inhibitor, are: Gly260 (part of Loop-Alpha), Glu284, Arg306, Glu322, Tyr324, Asn481, Glu485. With the exception of Gly260, which is only close enough to the lysine site in chain B of LAM_H2, these residues are seen to be involved in hydrogen bonding in every chain with the lysine substrate present – including with the lysine moiety of the lysyl-AMP intermediate in LAM_H2 chain A. In chains with the inhibitor cordycepin bound adjacent to lysine, Tyr483 is also seen to form a hydrogen bond with lysine – potentially suggesting that the presence of cordycepin does result in a slight adjustment of lysine position, but not one that seems significant. This is generally consistent with the findings of the preceding study, where *T. congolense* LysRS1 structures were solved with only one chain occupied by lysine and with both chains occupied by lysine.

One major difference consistently mentioned through this report is the position of Gly260, which would hydrogen bond with lysine when the 'Loop-Alpha' is in conformation 1. Loop-Alpha is only observed to be in conformation 1 in the LAM_H2 chain B, which is the one chain of all those in this study that has only lysine bound. In three of the four chains of the structures previously studied, named 'Semi-Complexed' and 'Fully-Complexed', where lysine is bound alone without an adjacent ligand, Loop-Alpha is also in conformation 1. Evidently this loop is involved in a conformational change during enzyme activity that also affects the position of Loop-Beta, an external surface loop only really visualised – and only with weak density – when Loop-Alpha is in conformation 1. In this position, Loop-Beta occupies much of the space where Loop-Alpha resides in conformation 2, making it highly likely that the movement is linked and that Loop-Beta becomes more stable, therefore better defined in a crystal structure, when Loop-Alpha is in that conformation 1.

In the one chain of the 'Semi-Complexed' structure where no ligands are bound, there is a lack of definitive electron density for Loop-Alpha suggesting that is conformationally flexible - though in that structure it has been arranged in a manner that is similar to conformation 2. It was initially suggested that that the binding of lysine stabilises this loop into an area where it can hydrogen bond with lysine. In view of the data observed in these newer structures, an updated explanation may be that the binding of a second ligand causes the reorganisation of that loop into an alternate conformation. This could be to promote tRNA binding, and thus trigger the second stage of the reaction where lysine is transferred from the lysyl-AMP intermediate onto tRNA and the resulting AMP is ejected.

However, in the published human LysRS structure 3BJU, Loop-Alpha is seen in conformation 1 as well. This could be due to an evolutionary difference, as the human LysRS Loop-Alpha equivalent has a different sequence (PGGAVAKP compared to AGGAAARP). It could also be because the ATP in that structure has not yet been hydrolysed to AMP; this could drive the conformational

change of the loop. In the 6CHD structure, Loop-Alpha is also in conformation 1 and again, this could be because ATP has not actually been hydrolysed – this structure contains a lysyl-AMP intermediate analogue. It is difficult to draw a definitive conclusion using the data currently available, but this is an area that warrants further investigation in future studies.

Moving onto the ATP binding site, residues identified that form hydrogen bonds with the second ligand are Arg306, Asn315, Glu478, Ile479, Asn481. Those seen to hydrogen bond with cordycepin are Arg306, His314, Glu478, Ile479 and Arg537. Of these, it is clear that Glu478 and Ile479 – along with Phe318 – are the most significant residues in this site. The former two hydrogen bond (Ile479 via its main chain carbonyl) with the hydroxyl groups present on the ribose moiety of AMP, cordycepin and the lysyl-AMP intermediate in these structures. That said, the fact that cladosporin can bind to the active site without possessing any such hydroxyl groups – and thus without hydrogen bonding with those two residues – indicates that it is possible for a molecule to bind without involving them directly, through the many hydrophobic or VDW contacts between site and ligand, so long as the ligand matches the complementary shape of the site. The latter residue mentioned, Phe318, is critical to the binding of the second ligand. Every structure determined with ligands in this site shows them to be pi-stacked with Phe318. It helps to correctly orientate the ligand and, in doing so, assists in ensuring that the ligand is positioned such that it can interact with other catalytic residues. Table 24 details the most common hydrophobic or VDW contacts in the (occupied) ATP-binding sites of the new structures resulting from this investigation, as reported by LigPlot. Note that residues already identified as involved in hydrogen bonding are excluded with the exception of His314, due to the lack of clarity surrounding the one instance in which it is reported (only by PyMOL) to form a hydrogen bond with cordycepin.

Residue involved in hydrophobic or VDW contact(s) and not in hydrogen bonding	Observed in the following sites
Phe318	All, including 4YCU and 3BJU
Gly534	All, including 4YCU and 3BJU
His314	All, except LAM_G4_A, including 4YCU and 3BJU
Thr313	LCM_F3_A, LAM_G4_B, LAM_H2_A and 3BJU

Table 24: Residues most commonly involved in hydrophobic or VDW contacts. This figure shows the residues at the ATP-binding site most often involved in hydrophobic or VDW contacts in the structures produced in this report. It is also stated whether those residues also form these contacts in the mammalian structures 3BJU and 4YCU. Residues already identified as forming hydrogen bonds with the ligands are not included.

Table 24 indicates that residues Phe318 and Gly534 are very important to ligand binding through hydrophobic or VDW contacts, with His314 being of similar importance (though only capable of VDW contacts, formed by the non-polar areas of the sidechain with the ligand).

Cordycepin has been identified as occupying the active sites of both chains in the LCM_F3 structure and so, as mentioned earlier, can be considered to act at least in some capacity as a competitive inhibitor to this system. The SASA difference calculations seen in Table 21 provide some initial supporting information in that cordycepin binding results in buried SASA very similar to that of cladosporin. Buried solvent-accessible surface area may be linked to partial closure of the otherwise open active site, which would be reflective of competitive inhibition. As cladosporin is another nucleoside mimic, already shown to be capable of targeting the LysRS of another fly-transmitted unicellular protozoan parasite (*Plasmodium falciparum*), this is a good sign.

A major difference between the trypanosome and mammalian enzymes is the position of Glu308 (325 in the human sequence). This is more unequivocal as it is easy to see that Glu308 is in a completely different orientation than its mammalian equivalent, twisted in the opposite direction and thus far from the active site in every single *T. congolense* LysRS1 structure but pointing toward the ATP-binding site and hydrogen bonding with the second ligand in every single mammalian structure examined. This means that there is increased volume in the parasitic enzyme ATP-binding site, as seen in the surface map view of this area in Figure 68 earlier, and ultimately this is the most obvious feature for exploitation when thinking of drug specificity. If a nucleoside analogue had a slightly larger group on the purine ring, perhaps a methyl or even an ethyl group, it may fit into the parasitic binding site whereas Glu325 will obstruct it from fitting into the mammalian binding site.

However, as seen in the study on cladosporin specificity (Fang et al., 2015), it is not only the most obvious residues that can provide specificity. In that study it was shown that in mammalian aaRSes three residues near to the active site but not directly involved in enzyme activity were responsible for the specificity cladosporin shows between and within aaRS families.

Generally considered to be important for stabilisation of the phosphate groups in aminoacyl-tRNA synthetases, the metal ions in these structures have proven continually elusive. Within the active sites themselves, in no structure solved in this study has there been any density that resembles a magnesium ion but elsewhere in the structures – without consistency – there has been spherical density that has been successfully interpreted by the placement of these ions. The lack of magnesium ions in the active site could be a result of the lack of phosphate groups, in the structures LAM_G4 and LAM_H2 at least, suggesting that they may have already moved out of the

active sites with the hydrolysed inorganic pyrophosphate. In the structure 1E1T, the inorganic pyrophosphate can be seen slightly askew of the lysyl-AMP intermediate with magnesium ions beside it, presumably about to be ejected from the site. As for LCM_F3, it could be simply that magnesium is not required for the successful binding of cordycepin to the enzyme – cordycepin has no phosphate groups, after all. One potential method to home in on where the magnesium ions are located is described in the 1E1T publication (Onesti, 2000), where magnesium ions were substituted with electron-dense manganese ions in a crystallography experiment.

Future directions

There are various directions that this research can take in the future. Firstly, with regard to the LysRS1 explored in this study, other inhibitors such as cladribine (identified earlier and listed in Figure 6) could be trialled in crystallisation studies to explore how the differences in these molecules affects their binding to the site (if they bind at all). Inhibitor assays could be performed using cordycepin in order to determine its effectivity. If a conclusive ATP-bound LysRS1 structure is sought, using non-hydrolysable ATP will likely provide this quickly if the same experimental conditions are used. As mentioned previously, to accurately locate the magnesium ions a crystallography experiment could be run where the magnesium ions are substituted with manganese ions.

Other aminoacyl-tRNA synthetase enzymes in *T. congolense* may also yield interesting results. It could be that there is a different aaRS that has even more significant differences between the parasitic and mammalian counterparts. Although time restraints prevented optimisation and diffraction analysis, as part of this study crystals have also been grown of *T. congolense* glycyl-tRNA synthetase. These displayed a very different morphology to those of the lysyl-tRNA synthetase, and grew in well D8 on two different plates. One example is seen in Figure 69.

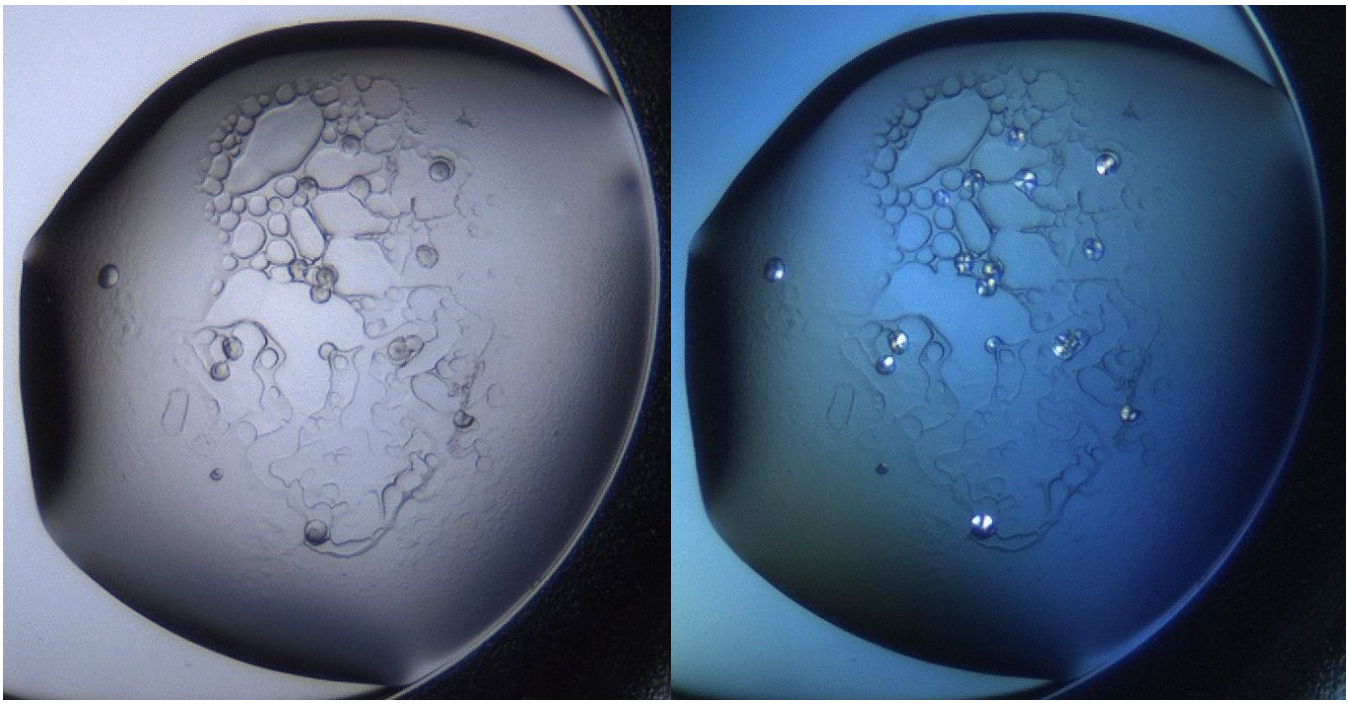

Figure 69: GlyRS + Glycine + ATP + $MgCl_2$ (GAM) crystals.
Sample solution consisted of 20mg/mL protein, 100mM glycine, 10mM ATP and 10mM $MgCl_2$. 0.5µl of this was added to 0.5µl well (D8, PACT Premier) solution – 20% PEG 6K, 0.2M NH_4Cl and 0.1M Tris pH8 buffer. Incubated at 4°C and inspected regularly.

For the purposes of determining specificity of any potential inhibitors and to allow for more appropriate comparisons in the future, the structure of bovine LysRS enzyme (or the equivalent of whichever other aaRS is being studied) would be valuable and would eliminate the need for other mammalian structures as approximations between the host and parasite structures. As mentioned earlier, the sequence of this enzyme is unreviewed and so even confirming that would allow sequence analyses to definitively determine, for example, whether there is an equivalent of Glu308. From then, hopefully it could be expressed and then crystallised with different ligands, though of course this is not necessarily an easy feat.

As climate change makes other areas of the world more habitable to these parasites, and globalisation aids their spread, the problem of Nagana is only going to worsen unless more pharmaceutical attention is paid to it. Hopefully the structures produced in this study, and the insights gained from them, will inform and inspire future interest in this underrepresented but important area.

Bibliography

Adams, P.D. et al. (2010) “PHENIX: a comprehensive Python-based system for macromolecular structure solution”, *Acta Crystallographica Section D Biological Crystallography*, 66(pt2), p213-221. doi: 10.1107/S0907444909052925

Altschul, S.F. et al. (1990) “Basic local alignment search tool.”, *Journal of Molecular Biology*, 215(3), p403-410. doi: 10.1016/S0022-2836(05)80360-2

Battenberg, O.A. et al. (2013) “Target profiling of 4-hydroxyderricin in S. aureus reveals seryl-tRNA synthetase binding and inhibition by covalent modification.”, *Molecular BioSystems*, 9(3), p343-51. doi: 10.1039/c2mb25446h

Battye, T.G. et al. (2011) “IMosflm: a new graphical interface for diffraction-image processing with MOSFLM”, *Acta Crystallographica Section D Biological Crystallography*, 67(pt4), p271-81. doi: 10.1107/S0907444910048675

Berman, H.M. et al. (2000) “The Protein Data Bank”, *Nucleic Acids Research*, 28(1), p235-242. doi: 10.1093/nar/28.1.235

Beuning, P.J. and Musier-Forsyth, K. (2000) “Hydrolytic editing by a class II aminoacyl-tRNA synthetase.”, *Proceedings of the National Academy of Sciences of the United States of America,* 97(16), p8916-20. doi: 10.1073/pnas.97.16.8916

Bragg, W.H. & Bragg, W.L. (1913) “The reflection of X-rays by crystals”, *Proceedings of the Royal Society of London*, 88, p428-438. doi: 10.1098/rspa.1913.0040

Chen, V.B. et al. (2010) “MolProbity: all-atom structure validation for macromolecular crystallography.”, *Acta Crystallographica Section D Biological Crystallography*, 66(pt1), p12-21. doi: 10.1107/S0907444909042073

DeLano, W.L. (2002) “The PyMOL User’s Manual”. DeLano Scientific, San Carlos, CA, USA.

Drenth, J. (2009) “Principles of Protein X-Ray Crystallography”. Springer, New York.

Eriani, G. et al. (1990) “Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs.”, *Nature*, 347(6289), p203-6. doi: 10.1038/347203a0

Eyford, B.A. et al. (2011) “Differential protein expression throughout the life cycle of *Trypanosoma congolense*, a major parasite of cattle in Africa.” *Molecular and Biochemical Parasitology*, 177(2), p116-125. doi: 10.1016/j.molbiopara.2011.02.009

Evans, P. & McCoy, A. (2008) “An introduction to molecular replacement” *Acta Crystallographica Section D Biological Crystallography*, 64(pt1), p1-10. doi: 10.1107/S0907444907051554

Fang, P. et al. (2015) “Structural Basis for Specific Inhibition of tRNA Synthetase by an ATP Competetive Inhibitor.”, *Cell Chemical Biology*, 18;22(6), p734-44. doi: 10.1016/j.chembiol.2015.05.007

Freist, W. (1995) “Threonyl-tRNA synthetase.”, *Biological Chemistry Hoppe-Seyler*, 376(4), p213-24.

Gasteiger, E. et al. (2005) “Protein Identification and Analysis Tools on the ExPASy Server”, *John M. Walker (eds): The Proteomics Protocols Handbook*, Humana Press, pp.571-607. doi: 10.1385/1-59259-890-0:571

Gibson, W. (2003) “Species concepts for trypanosomes: from morphological to molecular definitions?” *Kinestoplastid Biology and Disease*, 10(2). doi: 10.1186/1475-9292-2-10

Guo, M. et al. (2008) “Crystal structure of tetrameric form of human lysyl-tRNA synthetase: Implications for multisynthetase complex formation.”, *Proceedings of the Natural Academy of Sciences of the United States of America*, 105(7):2331-6. doi: 10.1073/pnas.0712072105

Hoepfer, D. et al. (2012) “Selective and Specific Inhibition of the *Plasmodium falciparum* Lysyl-tRNA Synthetase by the Fungal Secondary Metabolite Cladosporin”, *Cell Host Microbe,* 11(6):654-663. doi: 10.1016/j.chom.2012.04.015.

Holt, H.R. et al. (2016) “Assessment of animal African trypanosomiasis (AAT) vulnerability in cattle-owning communities of sub-Saharan Africa”, *Parasites & Vectors*, 9, p.53. doi: 10.1186/s13071-016-1336-5.

Horn, D. (2014) “Antigenic variation in African trypanosomes”, *Molecular and Biochemical Parasitology*, 195(2), p.123-139. doi: 10.1016/j.molbiopara.2014.05.001

Jones, T.W. & Dávila, A.M. (2001) “Trypanosoma vivax—out of Africa.”, *Trends in Parasitology*, 17(2), p.99-101. doi: 10.1016/S1471-4922(00)017773

Kalidas, S. et al. (2014) “Genetic validation of aminoacyl-tRNA synthetases as drug targets in trypanosome brucei.”, *Eukaryotic Cell*, 13(4), p504-16. doi: 10.1128/EC.00017-14

Langousis, G. & Hill, K.L. (2014) “Motility and more: the flagellum of Trypanosoma brucei”, *Nature Reviews Molecular Cell Biology*, 12(7), p.505-518. doi: 10.1038/nrmicro3274

Laskowski, R.A. and Swindells, M.B. (2011) “LigPlot+: multiple ligand-protein interaction diagrams for drug discovery.”, *Journal of Chemical Information and Modelling*, 51(10), p2778-86. doi: 10.1021/ci200227u

Lee, B. & Richards, F.M. (1971) “The interpretation of protein structures: Estimation of static accessibility”, *Journal of Molecular Biology*, 55(3), p379-400. doi: 10.1016/0022-2836(71)90324-X

Ling Joy Pang, Y. et al. (2014) “tRNA synthetase: tRNA Aminoacylation and beyond”, *Wiley Interdisciplinary Reviews: RNA*, 5(4), p.461-480. doi: 10.1002/wrna.1224

Lukes, J. et al. (2002) “Kinetoplast DNA network: evolution of an improbable structure.”, *Eukaryotic Cell*, 1(4), p495-502. doi: 10.1128/EC.1.4.495-502.2002

Matthews, K.R. (2005) “The developmental cell biology of *Trypanosoma brucei*”, *Journal of Cell Science*, 118(pt2), p283-290. doi: 10.1242/jcs.01649

McCready-Fallon, S. (2018) “A high-throughput approach to discovery and characterisation of potential drug targets from African trypanosomes.”, Biochemistry PhD thesis, University of Bristol, Bristol UK.

Michels, P.A. et al. (2006) “Metabolic functions of glycosomes in trypanosomatids.” *Biochimica et Biophysica Acta*, 1763(12), p1463-77. doi: 10.1016/j.bbamcr.2006.08.019

Morrison, L.J. et al. (2016) “Animal African Trypanosomiasis: Time to Increase Focus on Clinically Relevant Parasite and Host Species”, *Trends in Parasitology*, 32(8), p599-608. doi: 10.1016/j.pt.2016.04.012

Muhanguzi, D. et al. (2014) “The burden and spatial distribution of bovine African trypanosomes in small holder crop-livestock production systems in Tororo District, south-eastern Uganda”, *Parasites & Vectors*, 603(7). doi: 10.1186/s13071-014-0603-6

Nakama, T. et al. (2001) “Structural basis for the recognition of isoleucyl-adenylate and an antibiotic, mupirocin, by isoleucyl-tRNA synthetase.”, *Journal of Biological* Chemistry, 276(50), p47387-93. doi: 10.1074/jbc.M109089200

Onesti, S. et al. (2000) “Active site of lysyl-tRNA synthetase: structural studies of the adenylation reaction.”, *Biochemistry*, 39(29), p8418-25. doi: 10.1021/bi0006722

Orelle, C. et al. (2013) “Tools for characterizing bacterial protein synthesis inhibitors.”, *Antimicrobial Agents and Chemotherapy*, 57(12), p5994-6004. doi: 10.1128/AAC.01673-13

Pouplana, R. & Schimmel, P. (2001) “Two classes of tRNA synthetases suggested by sterically compatible dockings on tRNA acceptor stem.”, *Cell*, 104(2), p191-3. doi: 10.1016/S0092-8674(01)00204-5

Rhodes, G. (2000) “Crystallography made crystal clear: A guide for users of macromolecular models.” San Diego, Academic Press.

Rogers, M.J. et al. (1994) “Functional communication in the recognition of tRNA by *Escherichia coli* glutaminyl-tRNA synthetase.”, *Proceedings of the National Academy of Sciences of the United States of America*, 91(1), p291-5. doi: 10.1073/pnas.91.1.291

Rotureau, B. & Van Den Abbeele, J. (2013) “Through the dark continent: African trypanosome development in the tsetse fly.”, *Frontiers in Cellular and Infection Microbiology*, 3:53. doi: 10.3389/fcimb.2013.00053

Steverding, D. (2008) “The History of African trypanosomiasis”, *Parasites & Vectors*, 1, p.3. doi: 10.1186/1756-3305-1-3

Taylor, G. (2003) “The phase problem.”, *Acta Crystallographica Section D Biological Crystallography*, 59(pt 11), p1881-90. doi: 10.1107/S0907444903017815

The Food and Agricultural Organisation of the United Nations (2017) “A field guide for the diagnosis, treatment and prevention of African Animal Trypanosomiasis”. Available at: http://www.fao.org/docrep/006/x0413e/X0413E02.htm (Accessed: 18th March 2018)

Tubiana, M. (1996) “Wilhelm Conrad Rontgen and the discovery of X-rays.”, *Bulletin de L'Académie Nationale de Médecine*, 180(1), p97-108.

Wilson, D.N. (2009) “The A-Z of bacterial translation inibitors.”, *Critical Reviews in Biochemistry and Molecular Biology*, 44(6), p393-433. doi: 10.3109/10409230903307311

Wing, R. et al. (1980) “Crystal structure analysis of a complete turn of B-DNA”, *Nature*, 287, p755-758. doi: 10.1038/287755a0

Winn, M.D. et al. (2011) “Overview of the CCP4 suite and current developments.”, *Acta Crystallographica Section D Biological Crystallography*, 67(pt4), p235-42. doi: 10.1107/S0907444910045749

Wlodawer, A. et al. (2008) “Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures”, *FEBS Journal*, 275(1), p1-21. doi: 10.1111/j.1742-4658.2007.06178.x